Fundamentos de aprendizagem estatística, KNN e regressão linear

Aula 2

Magno Severino
PADS - Modelos Preditivos
07/04/2021

Objetivos de aprendizagem

Ao final dessa aula você deverá ser capaz de

- Definir um método para avaliar a performance de um modelo;
- Compreender formas de estimar o erro de teste.
- Analisar performance de um modelo no conjunto de treinamento.
- Analisar performance de um modelo no conjunto de teste.
- Conceituar o modelo KNN;
- Interpretar, ajustar e aplicar um modelo linear;

Na aula passada...

Erro redutível e erro irredutível

- Considere que seja observada uma variável quantitativa Y e p preditoras $X = (X_1, X_2, \ldots, X_p)$.

- Assumimos que existe uma relação entre essas medidas que pode ser escrita de forma geral como

$$Y = f(X) + \epsilon,$$

- Estimativa para f: $\hat{f}$.

- Considere uma $\hat{f}$ e um conjunto de preditoras X. Se fixarmos $\hat{f}$ e X, então

$$\begin{aligned} E(Y - \hat{Y})^2 &= E(f(X) + \epsilon - \hat{f}(X))^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{redutível}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{irredutível}}. \end{aligned}$$

Decomposição do erro de predição em viés e variância

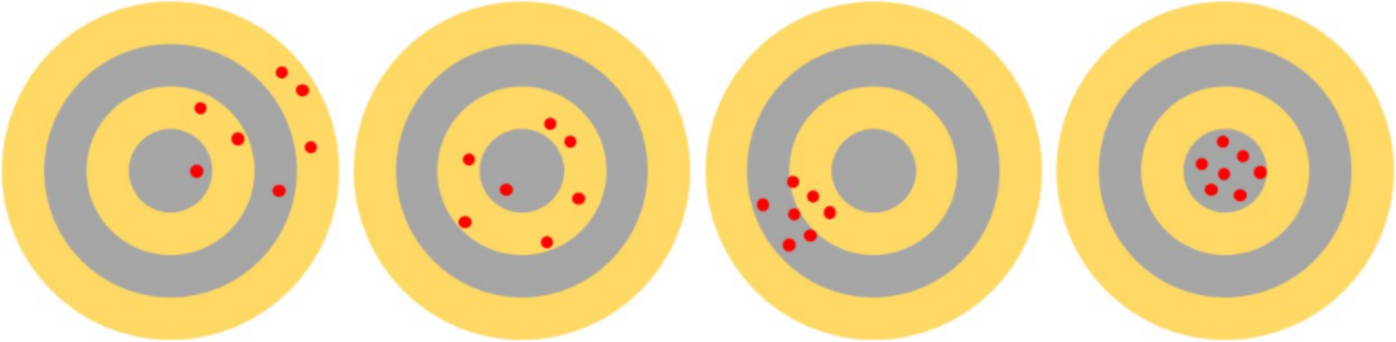
- Podemos decompor $E[(Y - \hat{f}(X))^2]$ em termos de viés e variância:

$$
\begin{aligned}
E[(Y - \hat{f}(x))^2] &= \int (\text{Vies}^2(\hat{f}(x)) + \text{Var}(\hat{f}(x)))dF_X(x) + \sigma^2 \\
&= \int \text{MSE}[\hat{f}(x)]dF_X(x) + \sigma^2.
\end{aligned}
$$

- O resultado acima nos diz que para minimizar o erro de predição esperado, temos que selecionar um método de aprendizagem estatística que tenha baixo viés e baixa variância simultaneamente.

Trade-off viés e variância

- $\text{Var}(\theta) = E(\theta^2) - E^2(\theta)$;
- $\text{Viés}(\theta) = E(\hat{\theta}) - \theta$;
- $\text{MSE}(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = \text{Viés}^2(\theta) + \text{Var}(\hat{\theta})$.



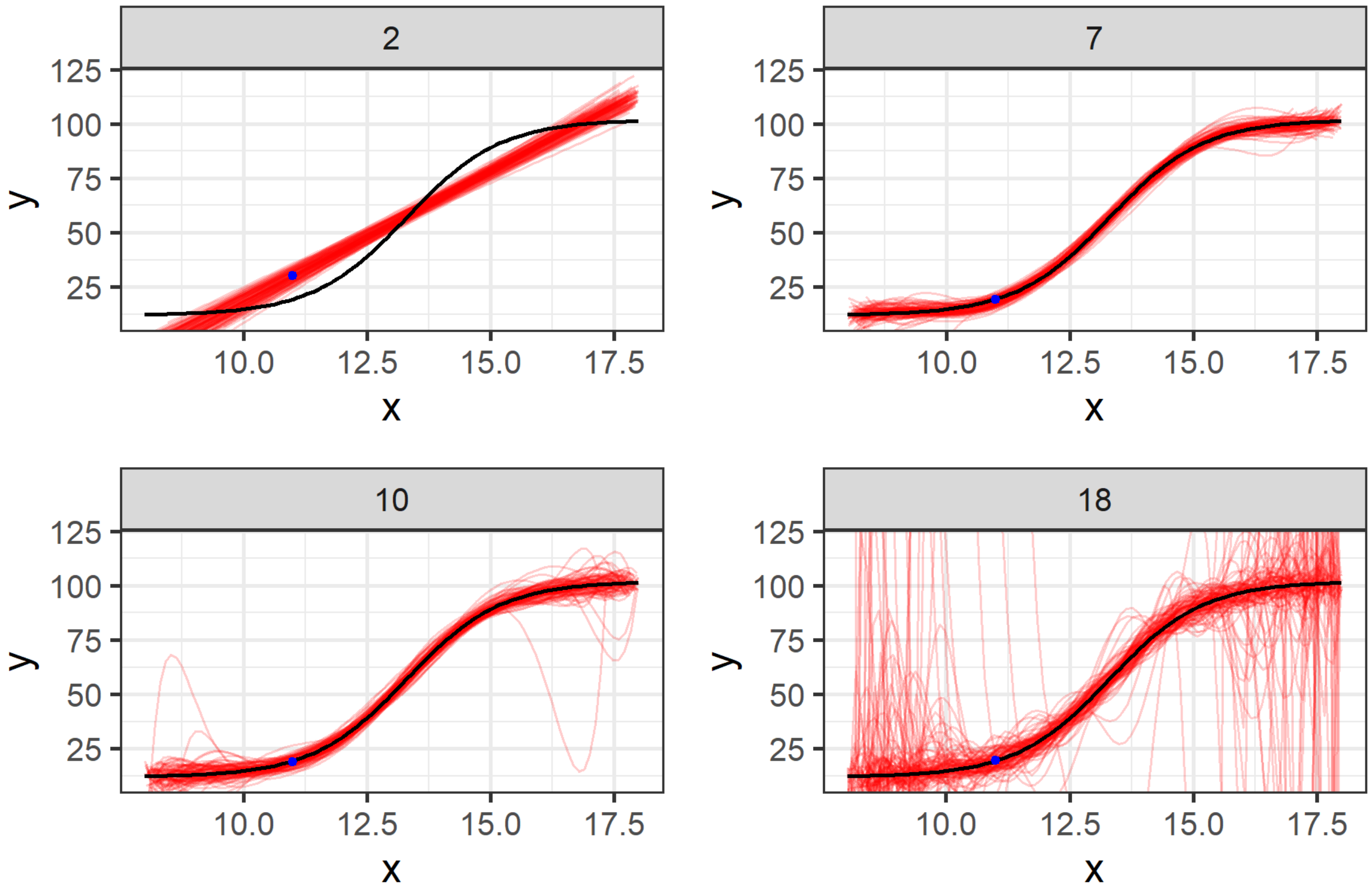
Simulação

- Estamos interessados em prever a renda anual Y de uma pessoa que possui x anos de escolaridade.
- Considere que a função que relaciona anos de escolaridade com a renda anual é

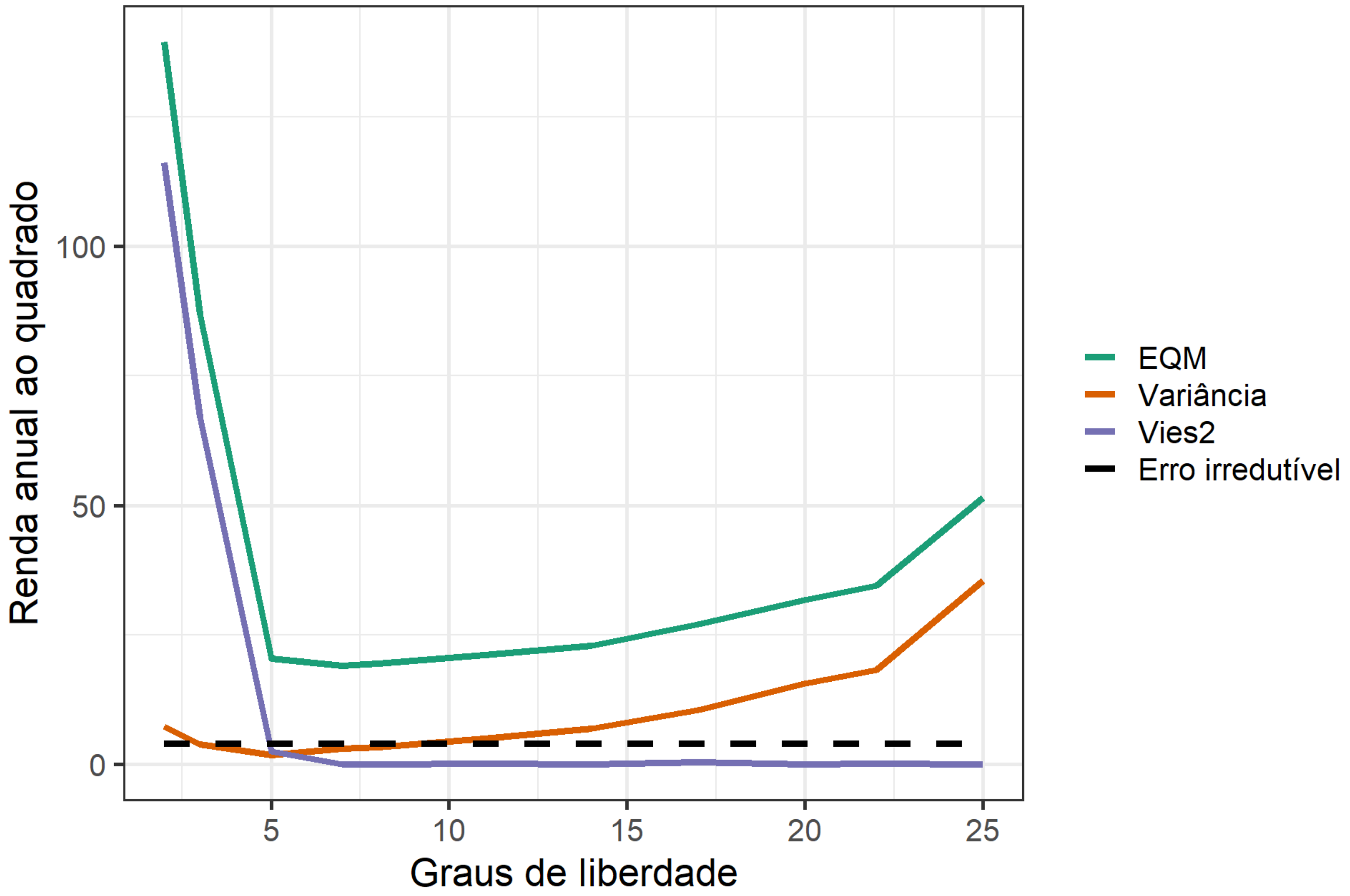
$$\begin{aligned} Y &= f(x) + \epsilon \\ &= 45 \times \tanh\left(\frac{x}{1{,}9} - 7\right) + 57 + \epsilon, \end{aligned}$$

em que $x \in [8, 18]$ e $\epsilon \sim N(0, 4^2)$.

Grafico resultados simulação



Trade-off viés e variância

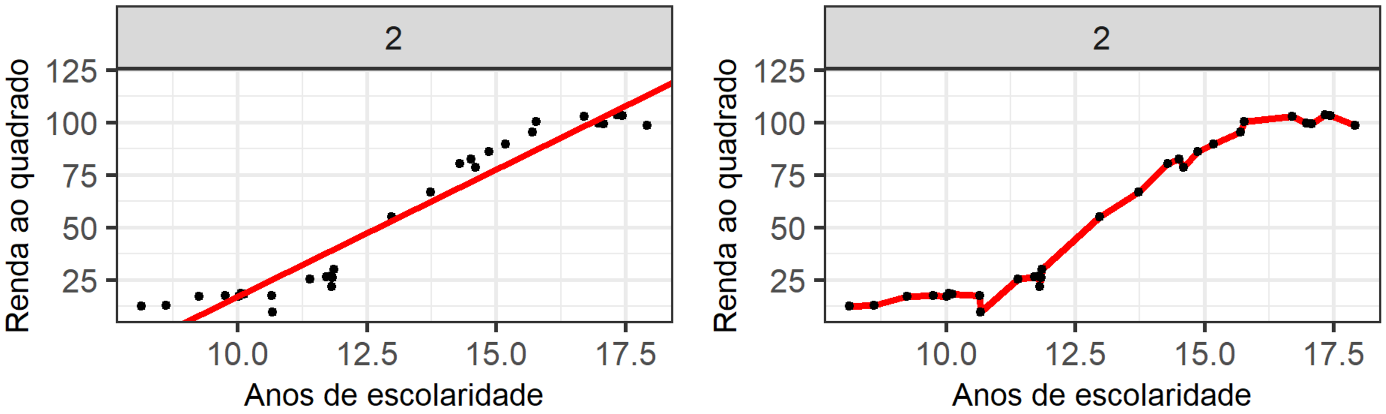


Problema

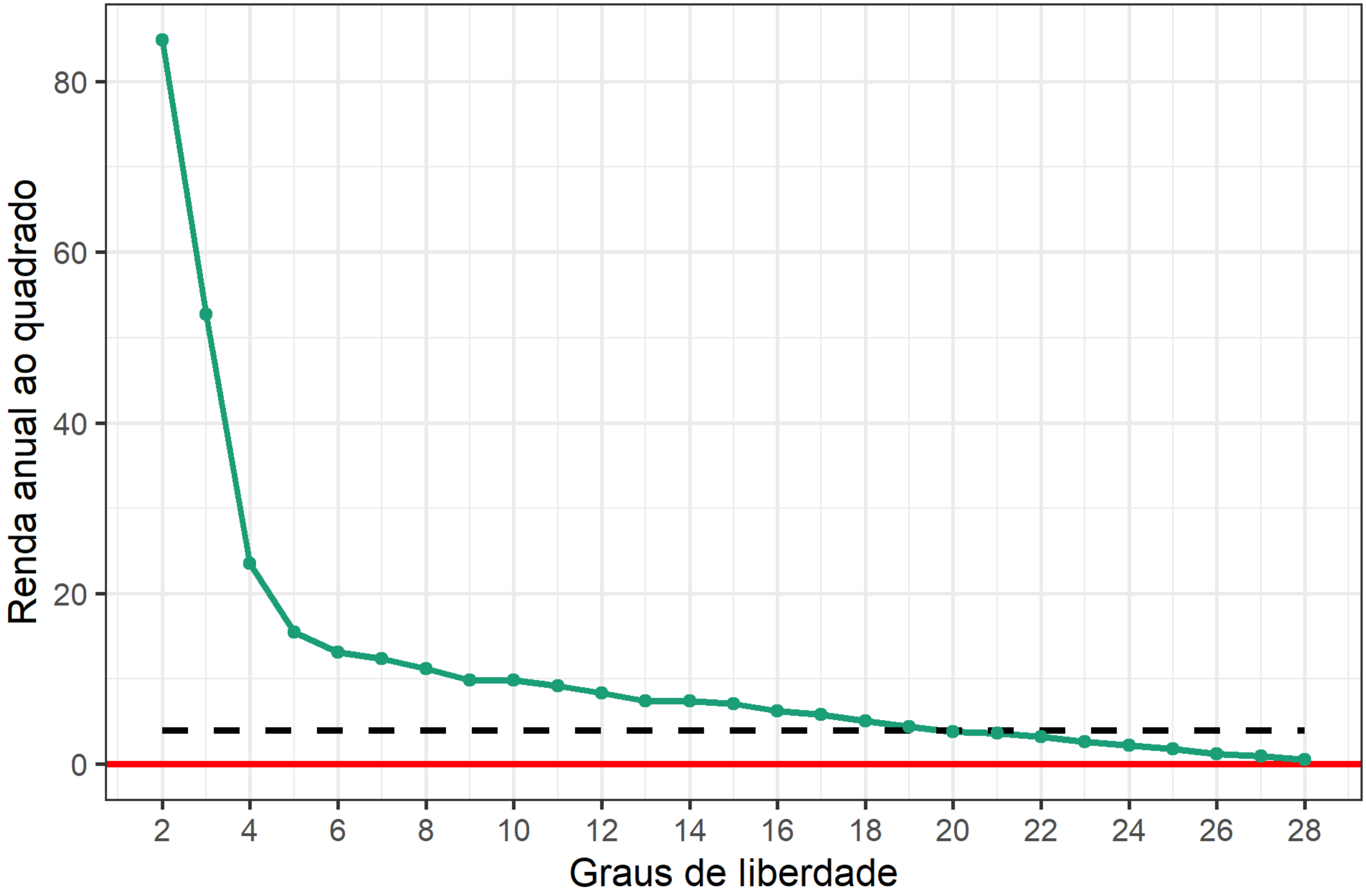
- Na prática, não conhecemos a função $f(\cdot)$!
- Como estimar o erro de predição $(f(x_i) - \hat{f}(x_i))^2$?
- Alternativa: considerar a diferença $(y_i - \hat{f}(x_i))^2$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2.$$

- Mas existe um problema!



Erro quadrático médio de treinamento



Prática R

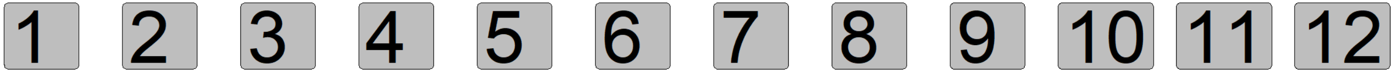
Como estimar o erro de predição?

- **MSE de teste**: calculado com dados que **não** pertencem ao conjunto de treinamento.
- Essa métrica é utilizada na seleção de modelos.
- Na prática não há um conjunto de teste disponível.
- Alternativas:
  - *Validation set approach*;
  - *Cross validation*;
  - *Bootstrap*.

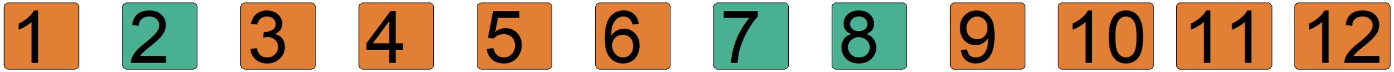
Validation set approach

Considere que o conjunto de dados tem 12 observações.

Ideia: considerar 75% dos dados para treinamento e 25% para validação.



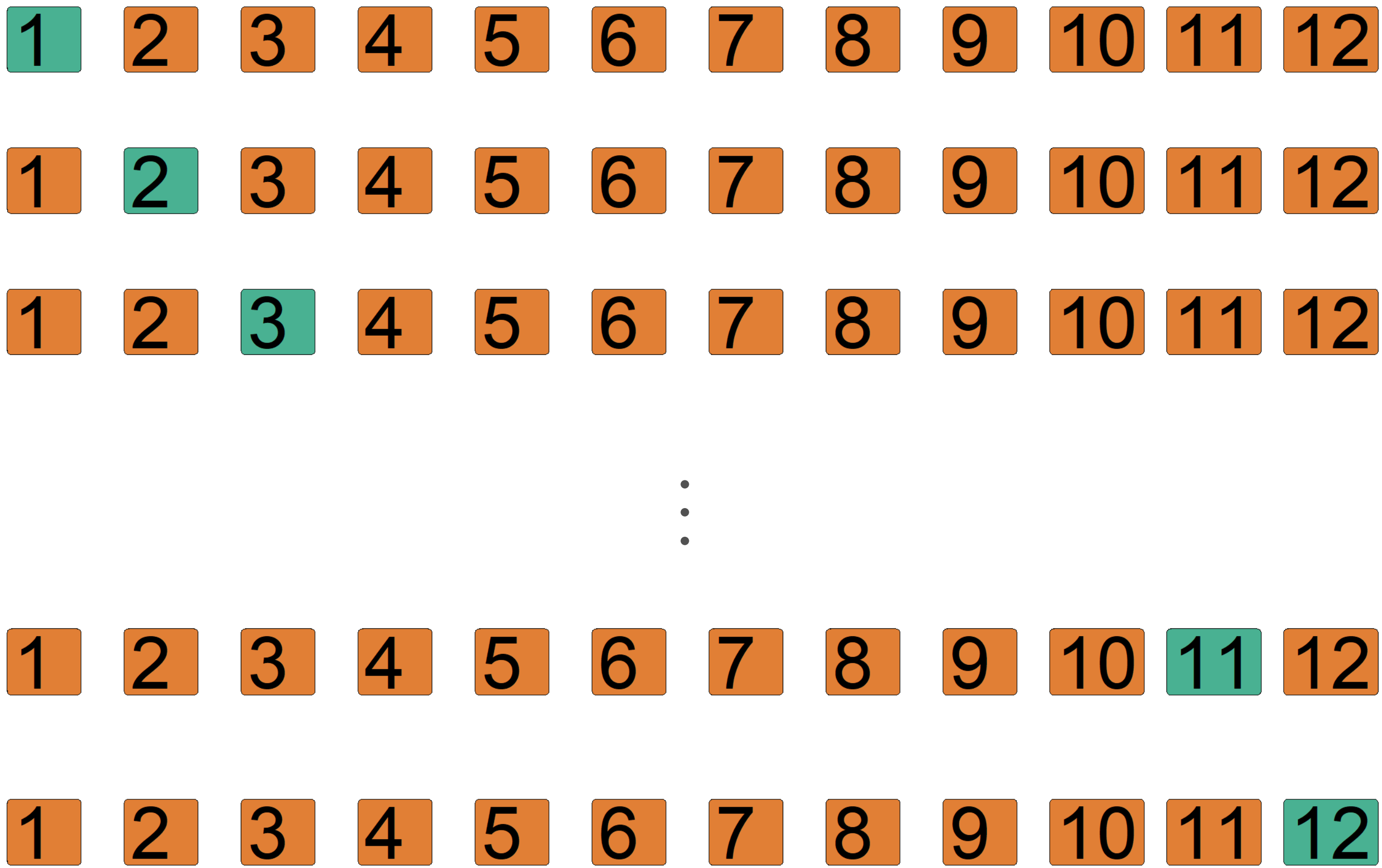
Veja abaixo uma separação possível.



Problema?

Leave-one-out cross-validation (LOOCV)

Nesta abordagem, vamos considerar como conjunto de validação uma observação por vez.



Problema?

k-fold cross-validation

Considere os seguintes dados.

1	2	3	4	5	6	7	8	9	10	11	12

Vamos dividi-los aleatoriamente em $k = 4$ lotes.

1	4	6	5	9	12	3	7	10	2	8	11

Assim, obtemos os seguintes conjuntos de treino e teste.

Treino 1	2	3	5	7	8	9	10	11	12	1	4	6
Treino 2	1	2	3	4	6	7	8	10	11	5	9	12
Treino 3	1	2	4	5	6	8	9	11	12	3	7	10
Treino 4	1	3	4	5	6	7	9	10	12	2	8	11

Relação entre as duas abordagens

- LOOCV é um caso particular da validação cruzada quando $k = n$.
- LOOCV é um procedimento que tem maior custo computacional.
- Em geral, LOOCV apresenta uma variância maior do que a validação cruzada em k lotes. Uma vez que

$$\mathrm{Var}(CV_{(n)}) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(\mathrm{MSE}_i) + \frac{2}{n^2}\sum_{1\leq i\leq j\leq n}\mathrm{Cov}(\mathrm{MSE}_i, \mathrm{MSE}_j).$$

- A prática indica que as escolhas $k = 5$ e $k = 10$ são um bom compromisso entre viés, variância e custo computacional.

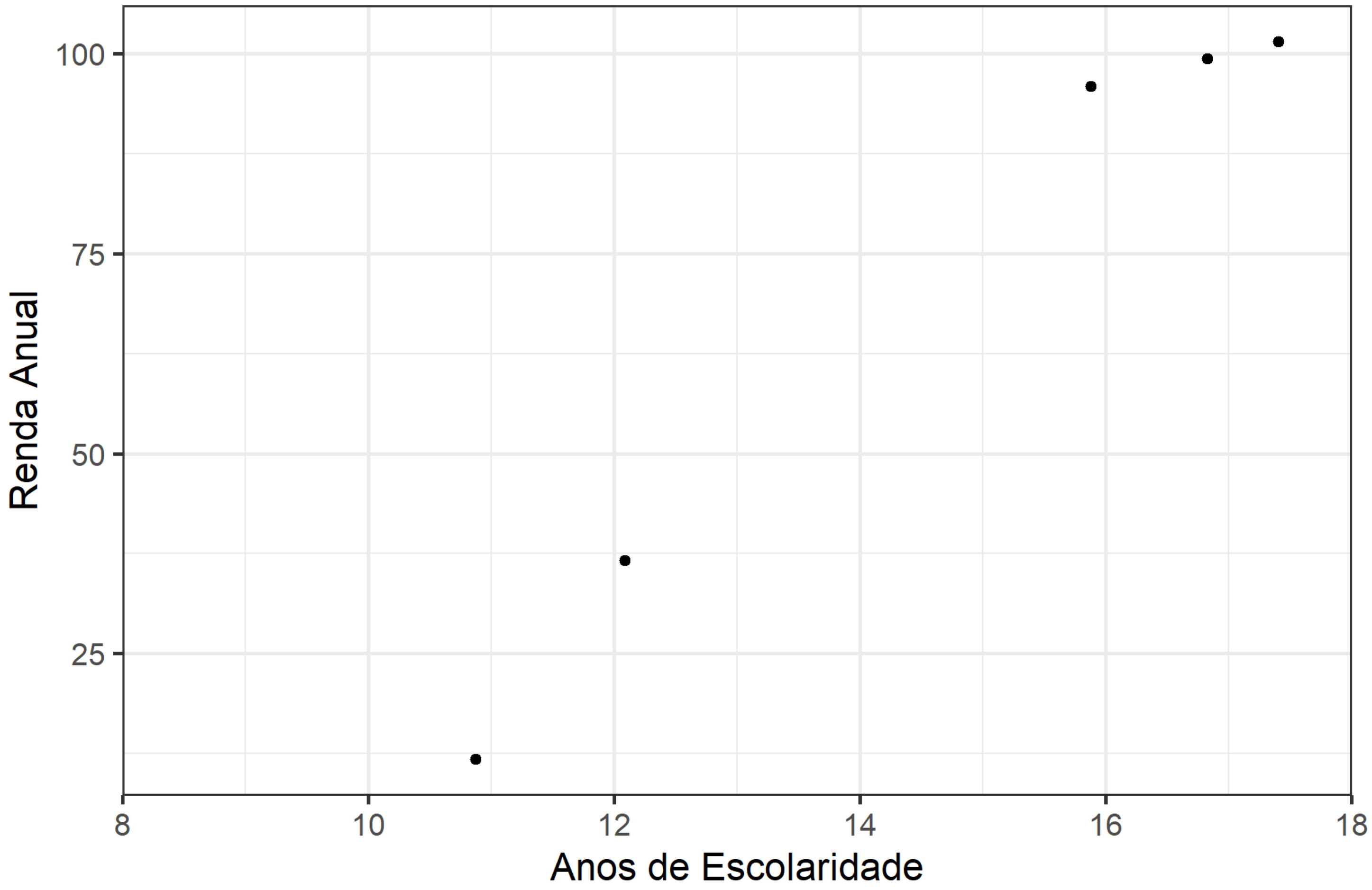
Prática *R*

- Objetivo: retomar os dados simulados e definir os graus de liberdade utilizando a validação cruzada.

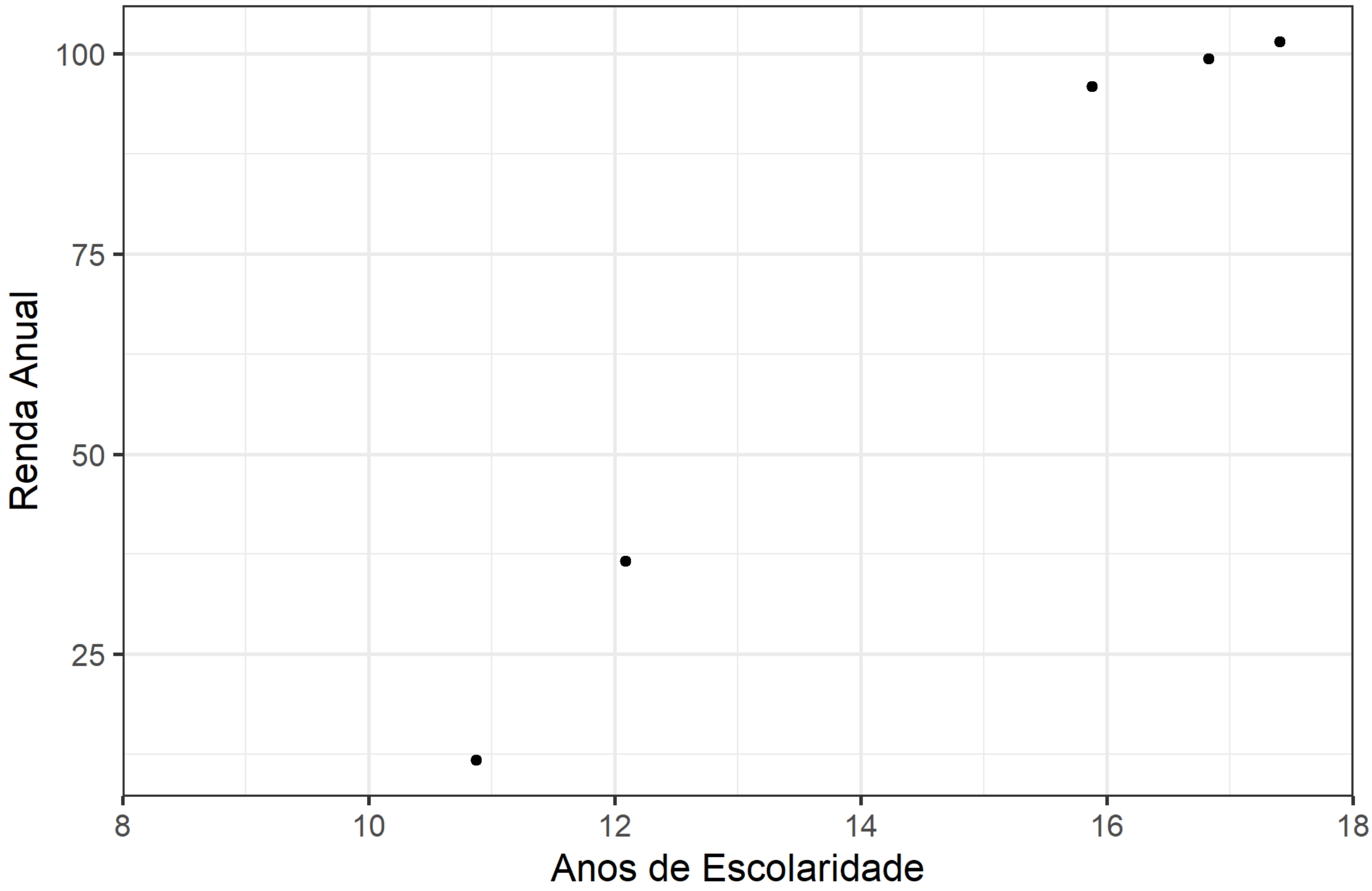
```
n_obs <- 30

set.seed(123)

dados <- tibble(x = sort(runif(n = n_obs, min = 8, max = 18)),
                y = 45*tanh(x/1.9 - 7) + 57 + rnorm(n = n_obs,
                                                    mean = 0,
                                                    sd = 4))

dados %>%
  ggplot(aes(x, y)) +
  geom_point() +
  theme_bw()

folds <- 5

lote <- sample(1:folds, size = n_obs, replace = TRUE)
```

Regressão *k nearest neighbours* (KNN)

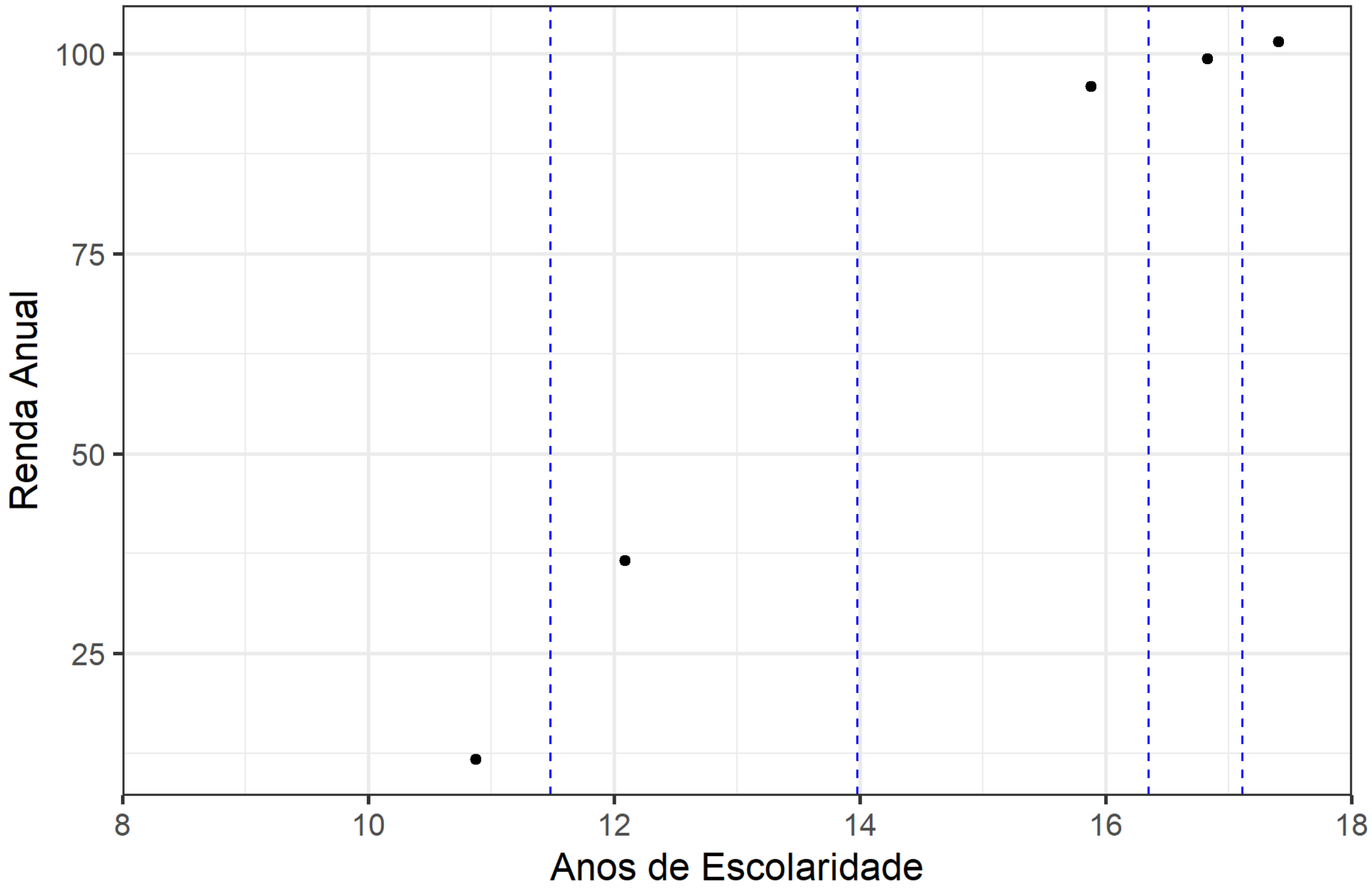
Regressão KNN



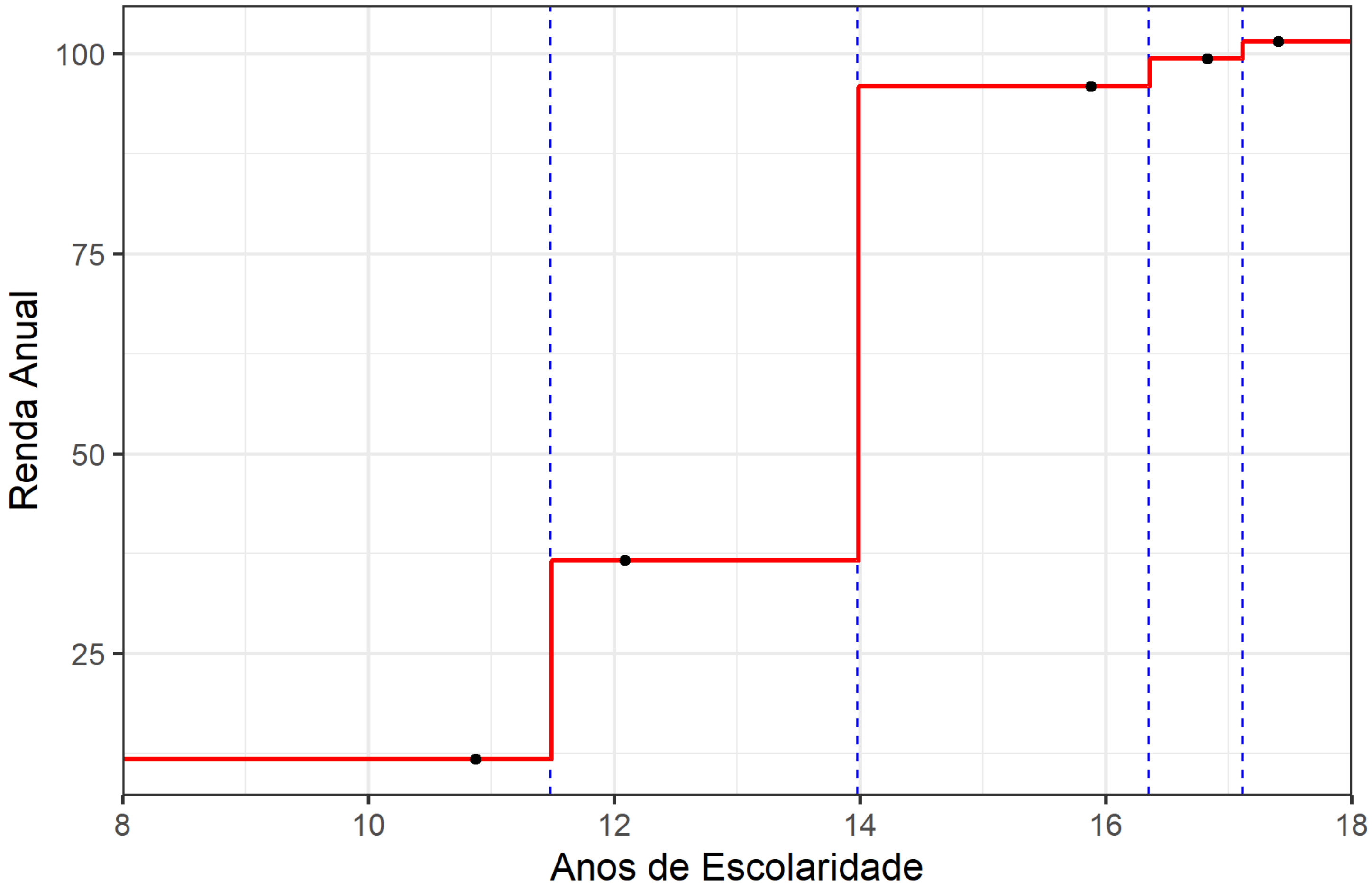
Regressão 1-NN



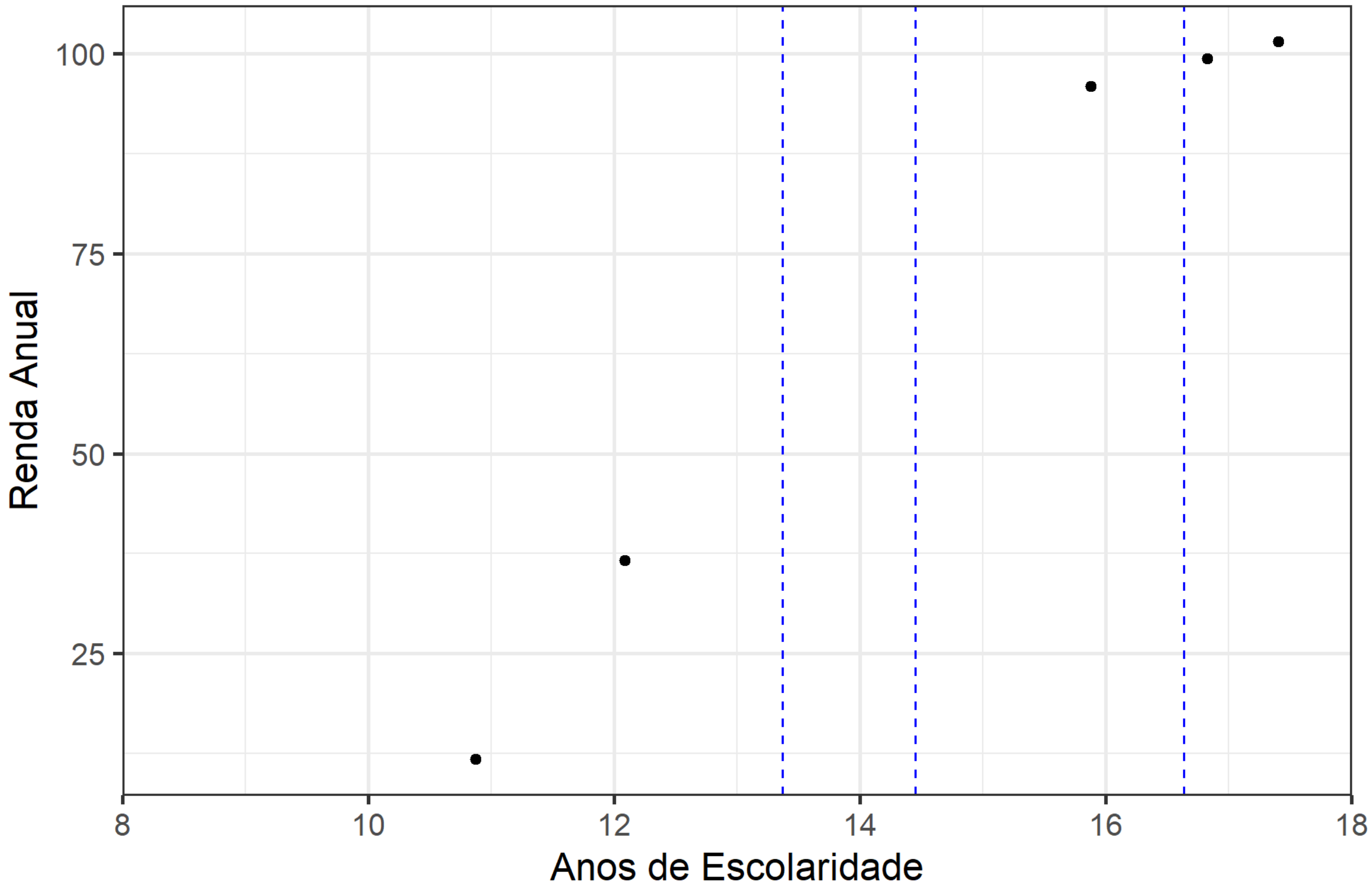
Regressão 1-NN



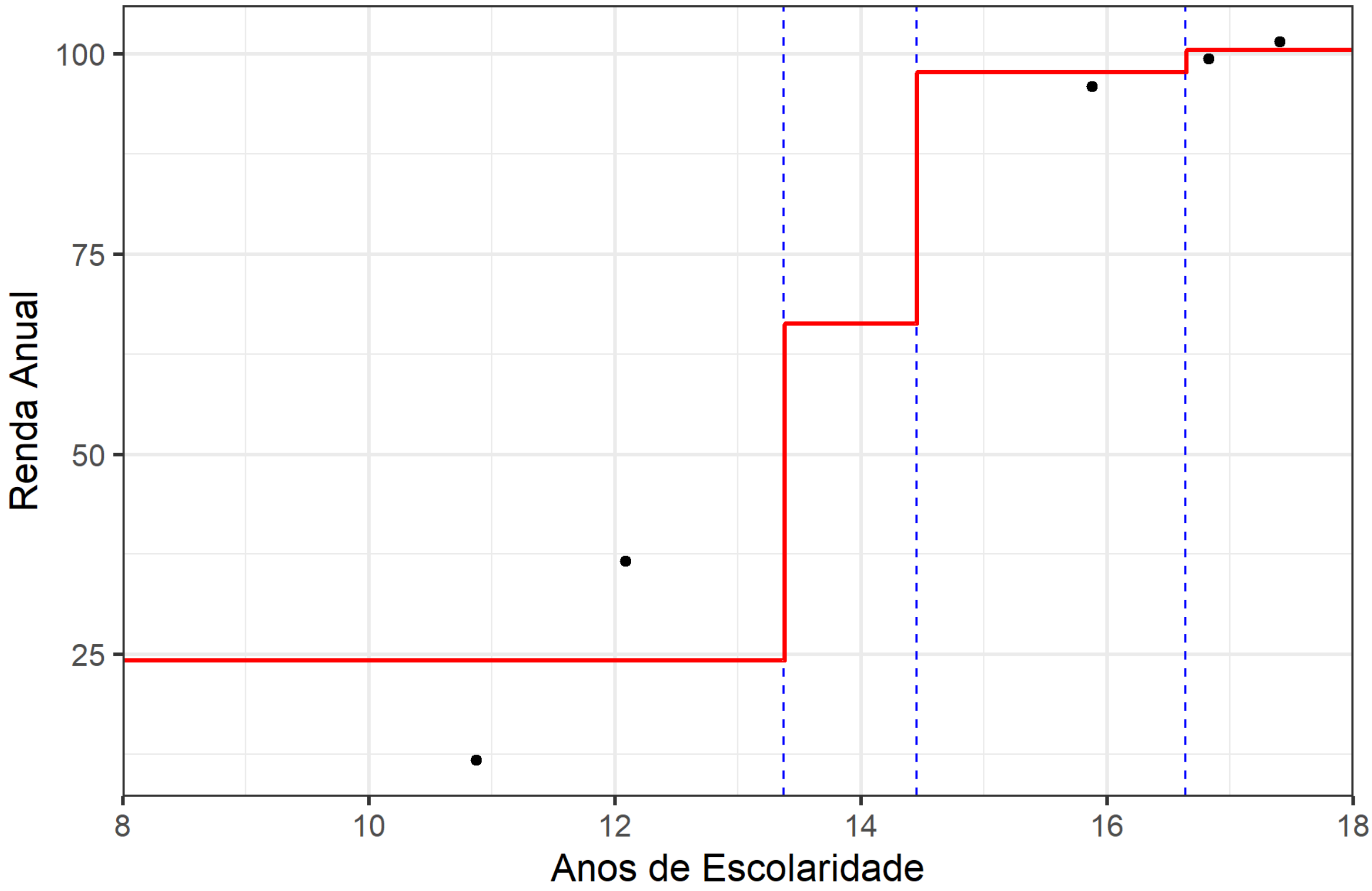
Regressão 1-NN



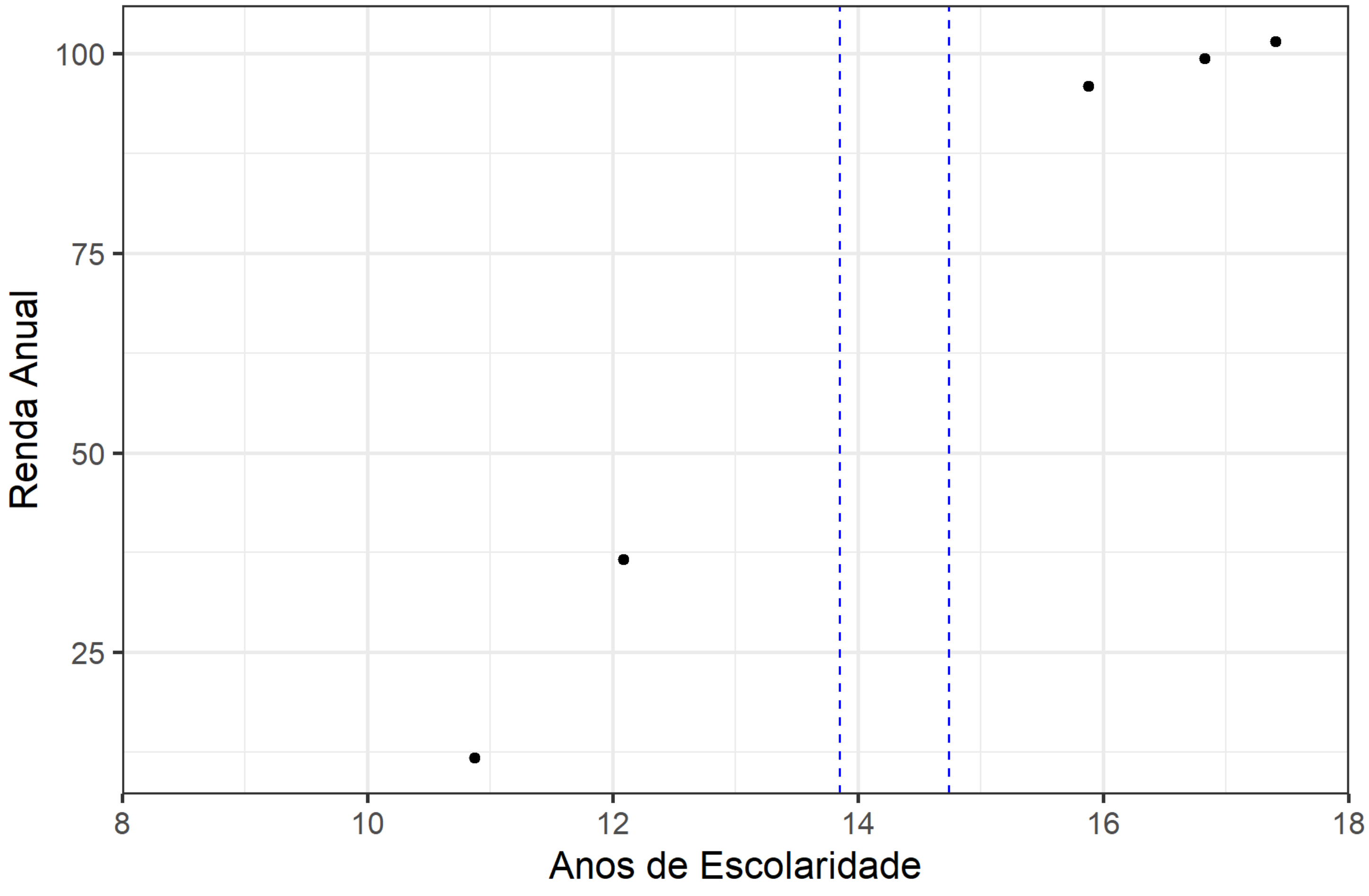
Regressão 2-NN



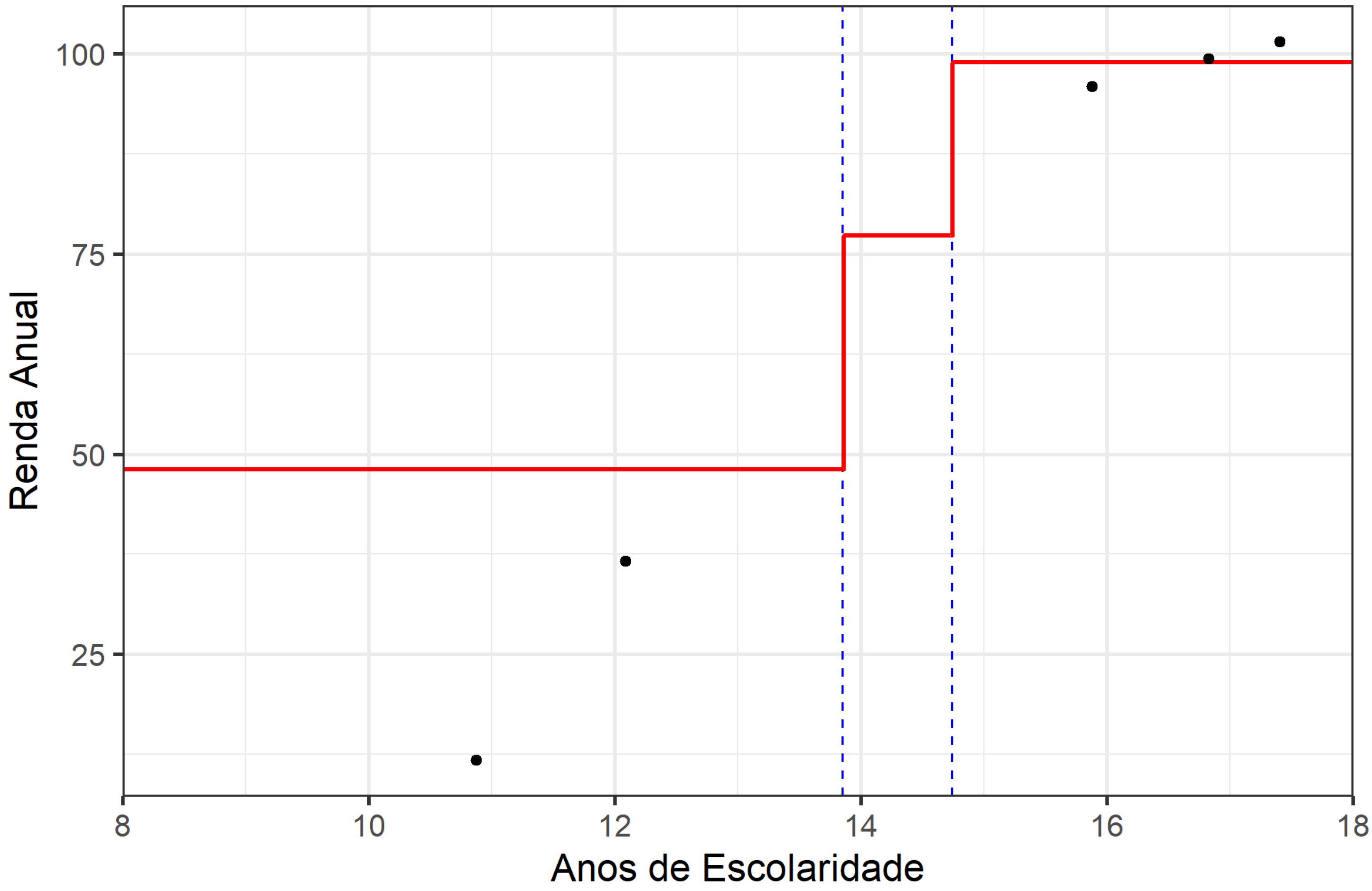
Regressão 2-NN



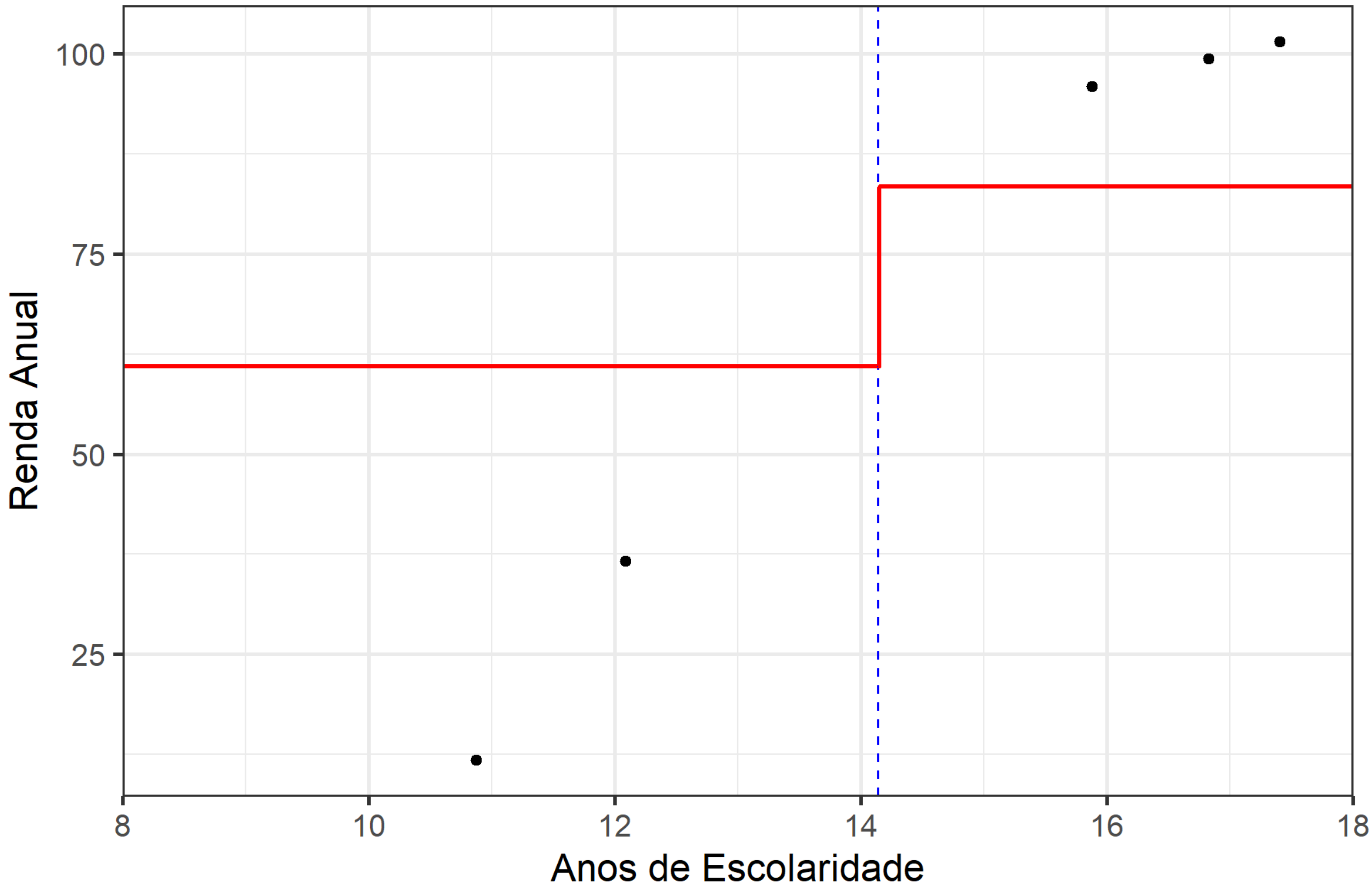
Regressão 3-NN



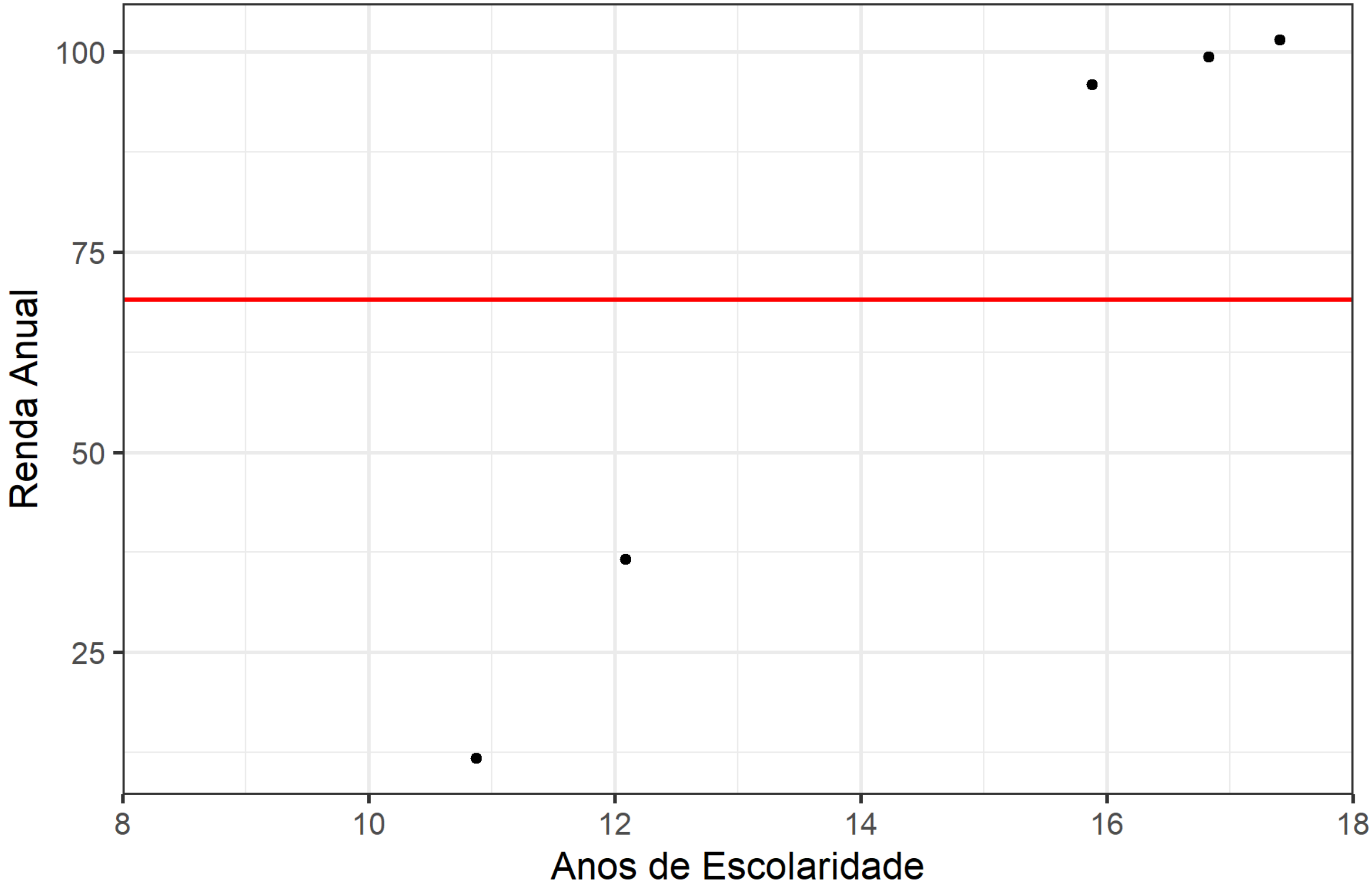
Regressão 3-NN



Regressão 4-NN



Regressão 5-NN



Prática no R

- Objetivo: retomar os dados simulados e ajustar o modelo KNN, definindo o número de vizinhos utilizando validação cruzada.

- Parâmetros da simulação:

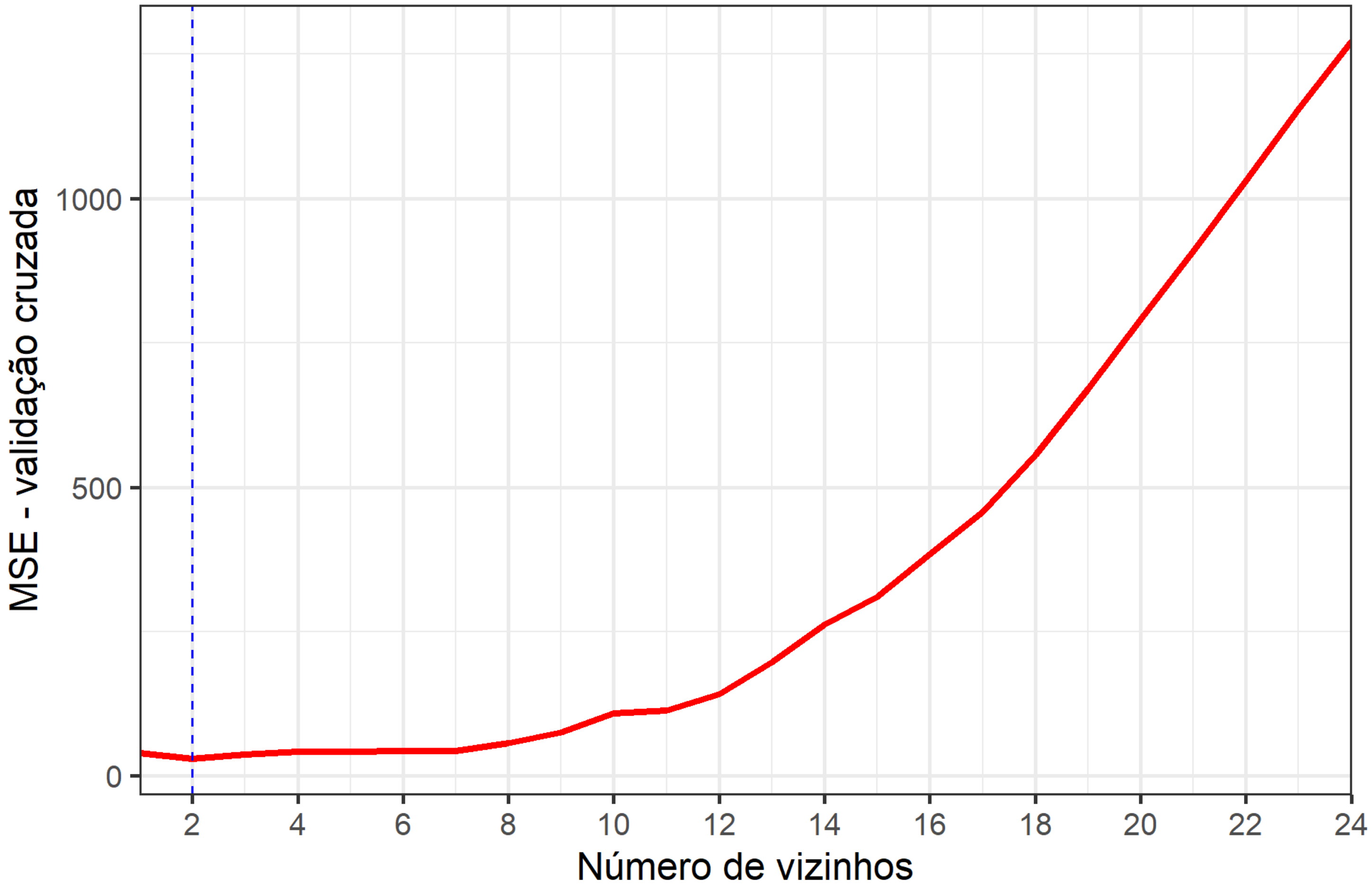
  - **n_sample**: número de observações do conjunto de treinamento (30);

  - **folds**: número de lotes para validação cruzada (5-fold);

  - **n_vizinhos**: um valor entre 2 e 30.

MSE validação cruzada 5 lotes

Considerando os dados simulados, com `n_sample = 30`.



Introdução

Regressão linear

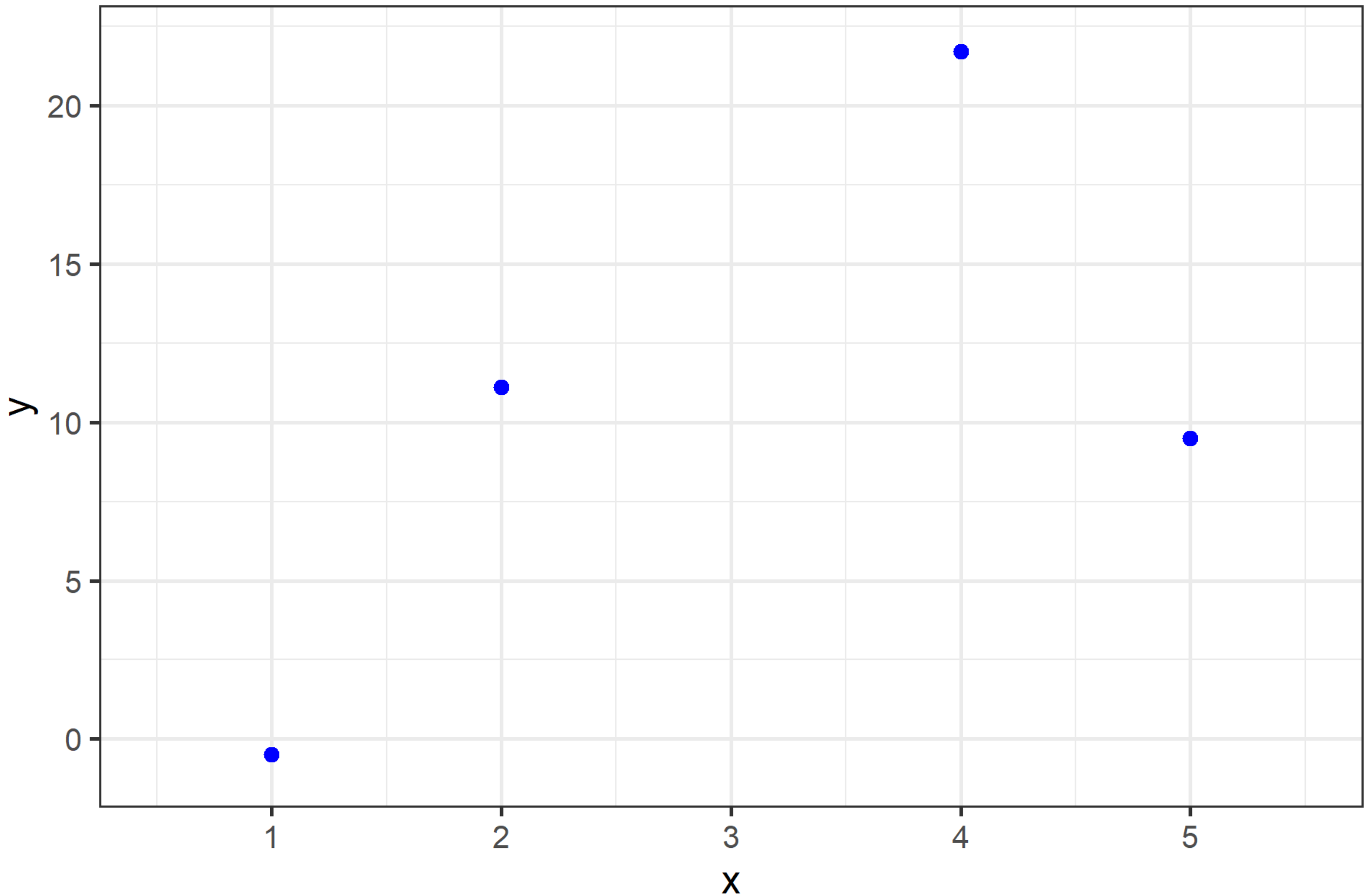
- Assumimos que existe uma relação aproximadamente linenar entre X e Y.
- Matematicamente, podemos escrever esta relação como

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

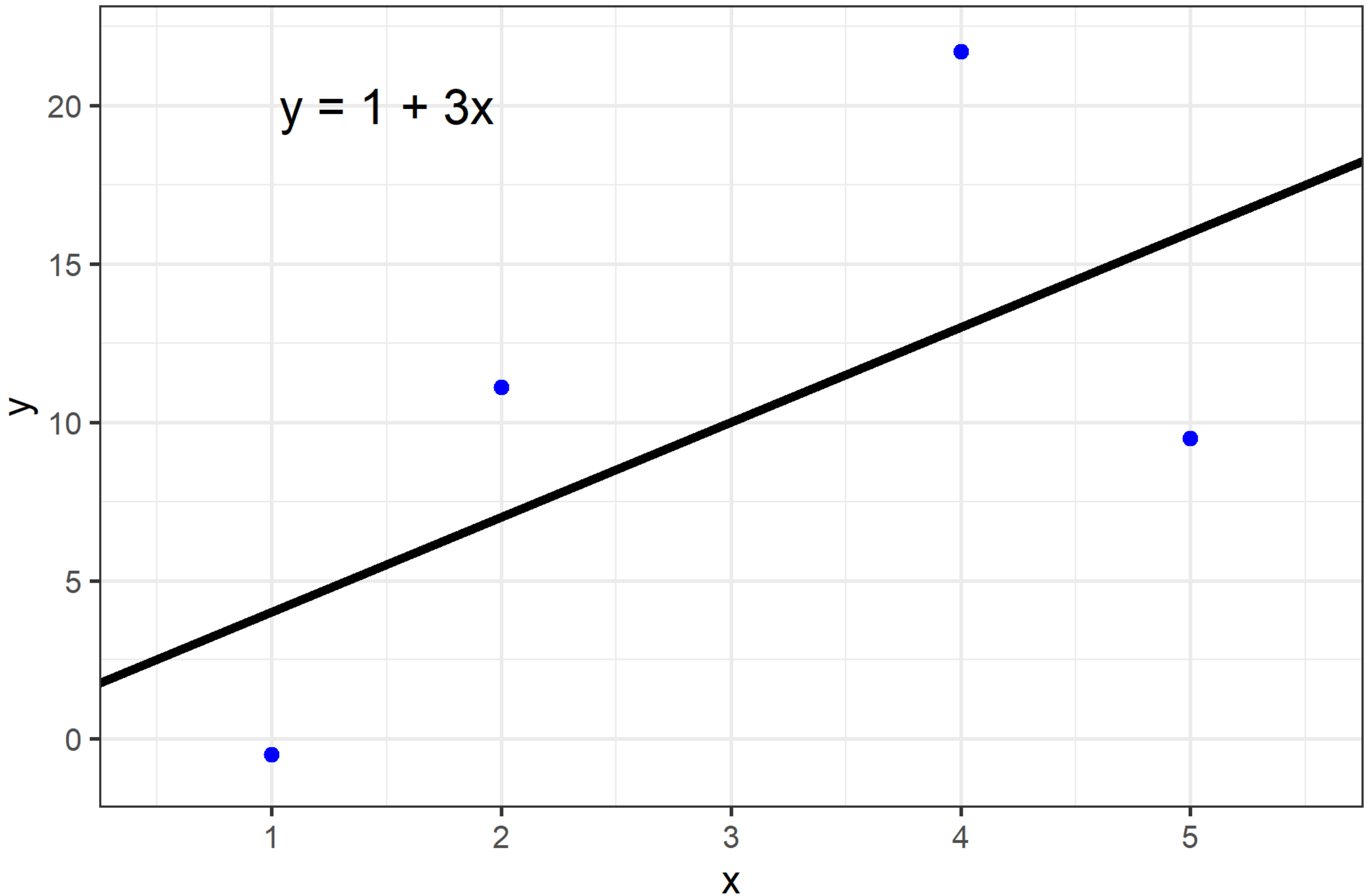
- Baseado no conjunto de treinamento, podemos obter estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$ para os coeficientes β_0 e β_1 do modelo e assim estimar o valor de Y quando $X = x$:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

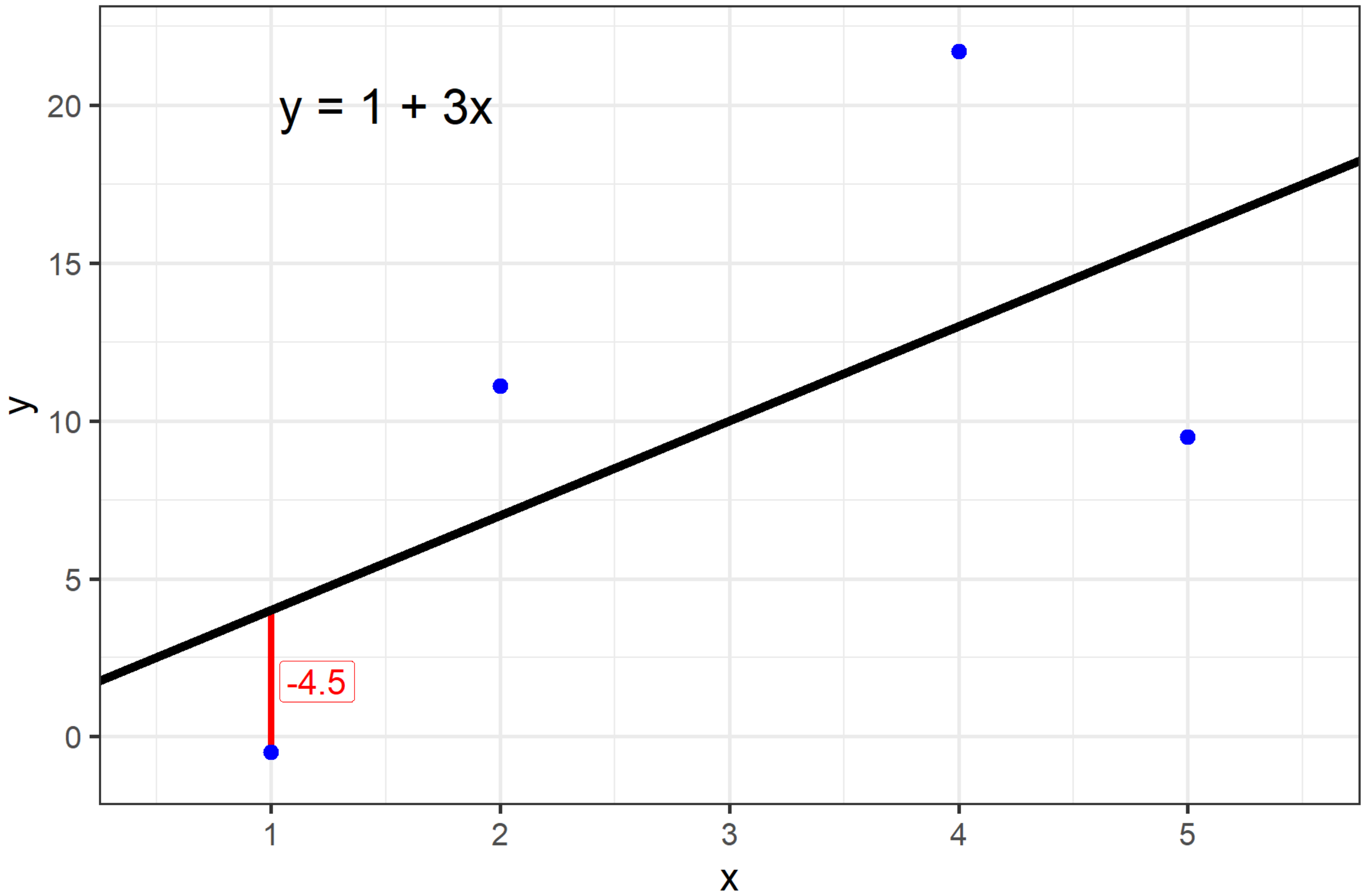
Regressão linear



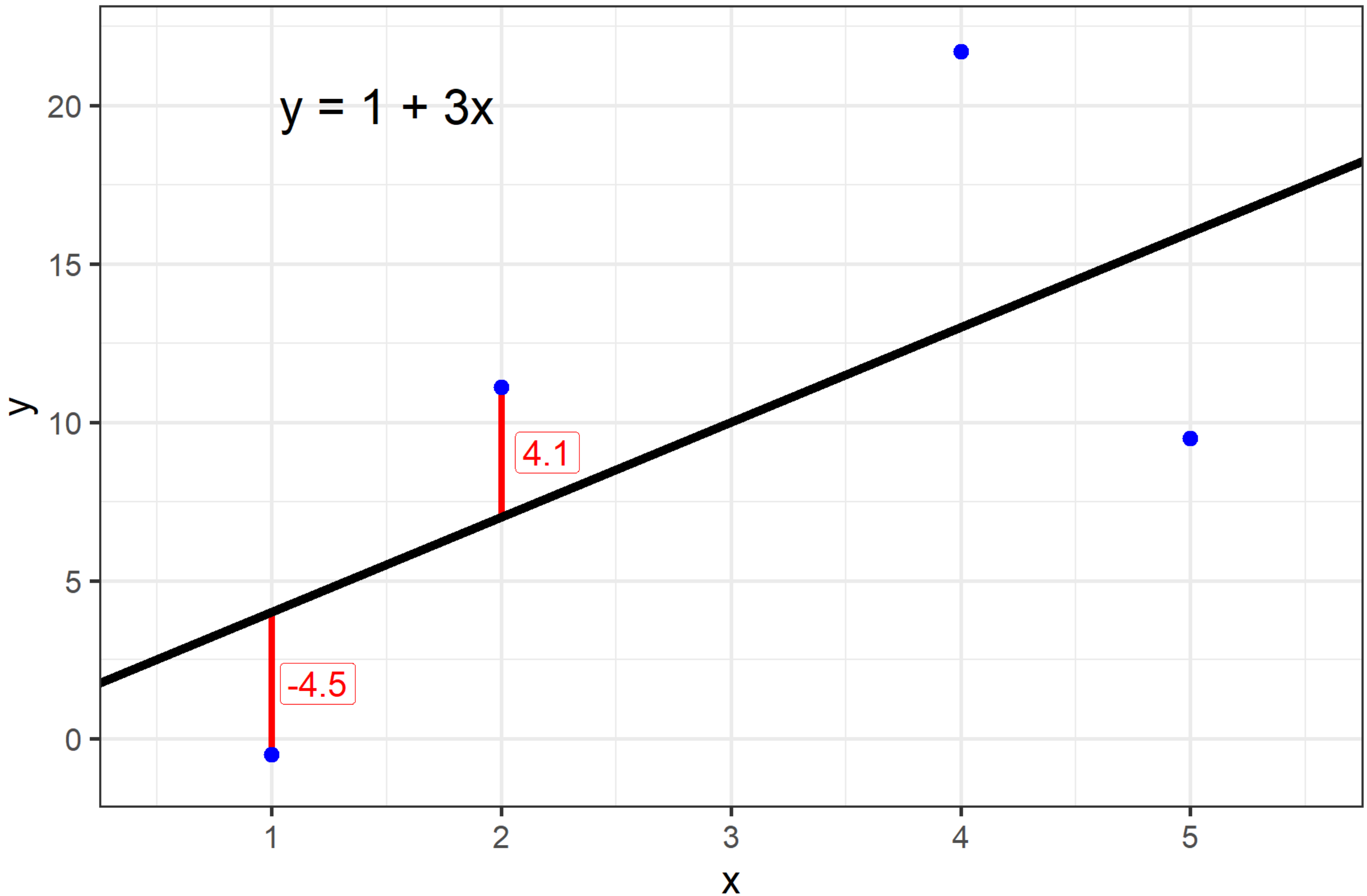
Regressão linear



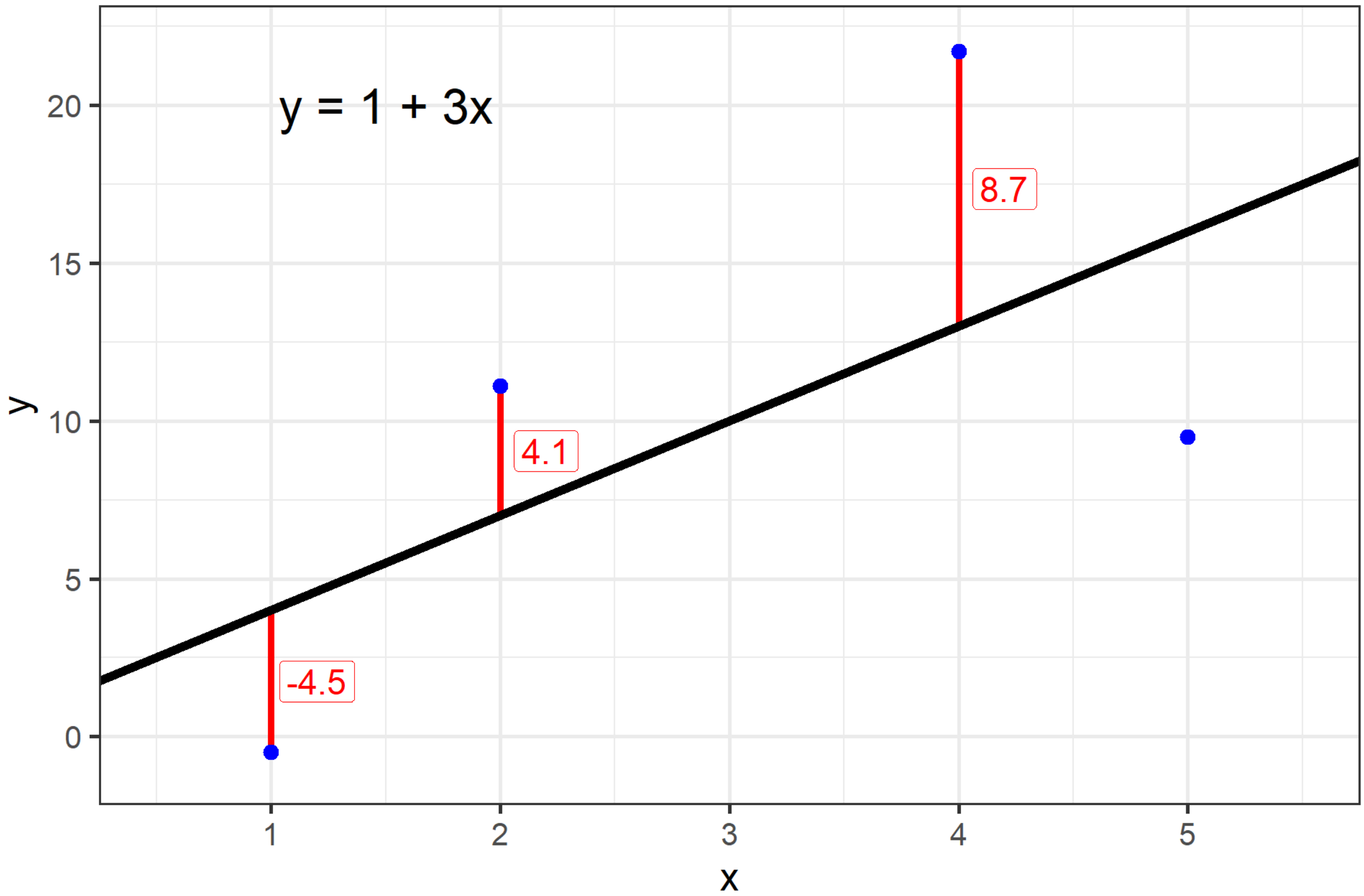
Regressão linear



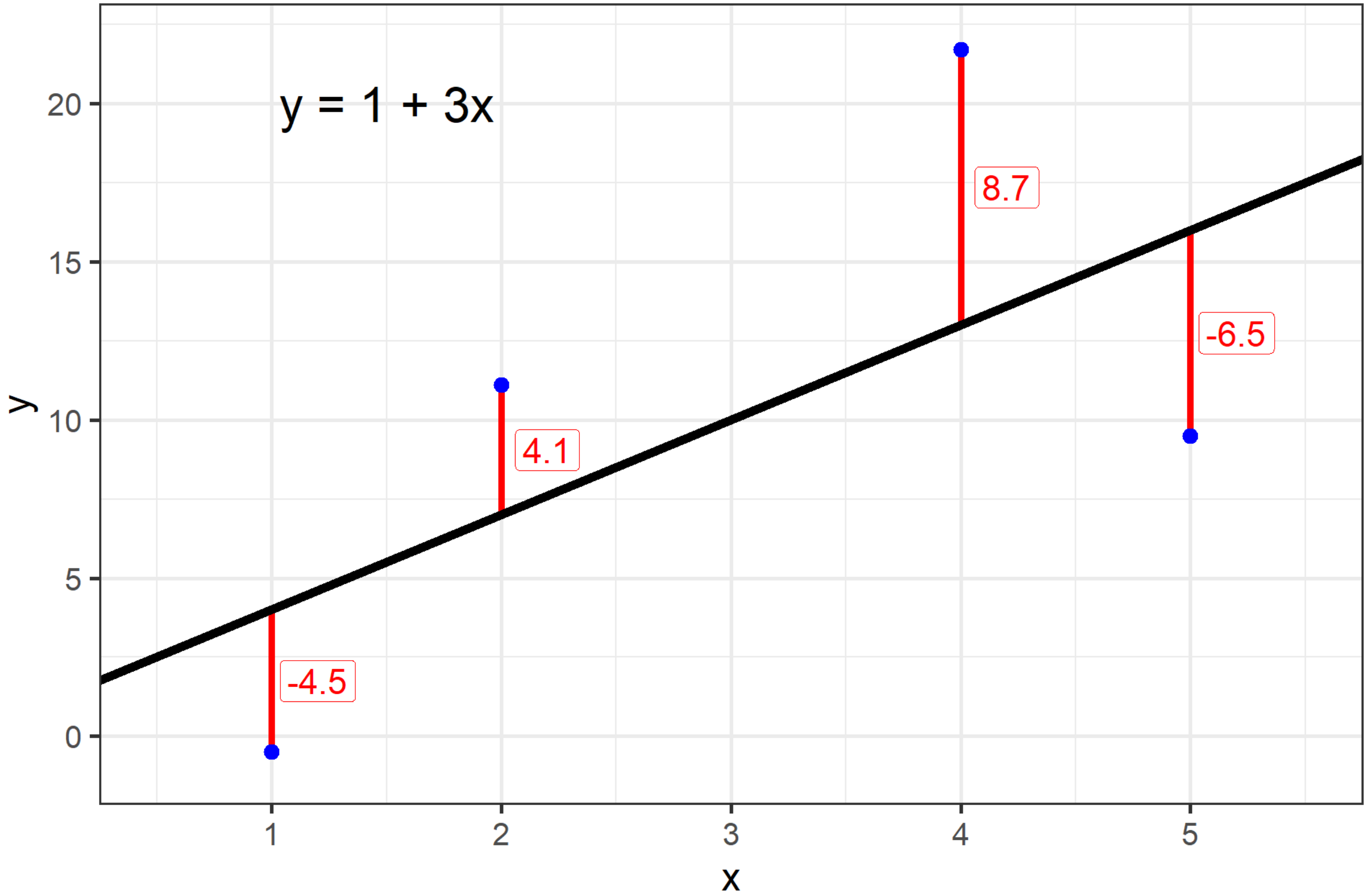
Regressão linear



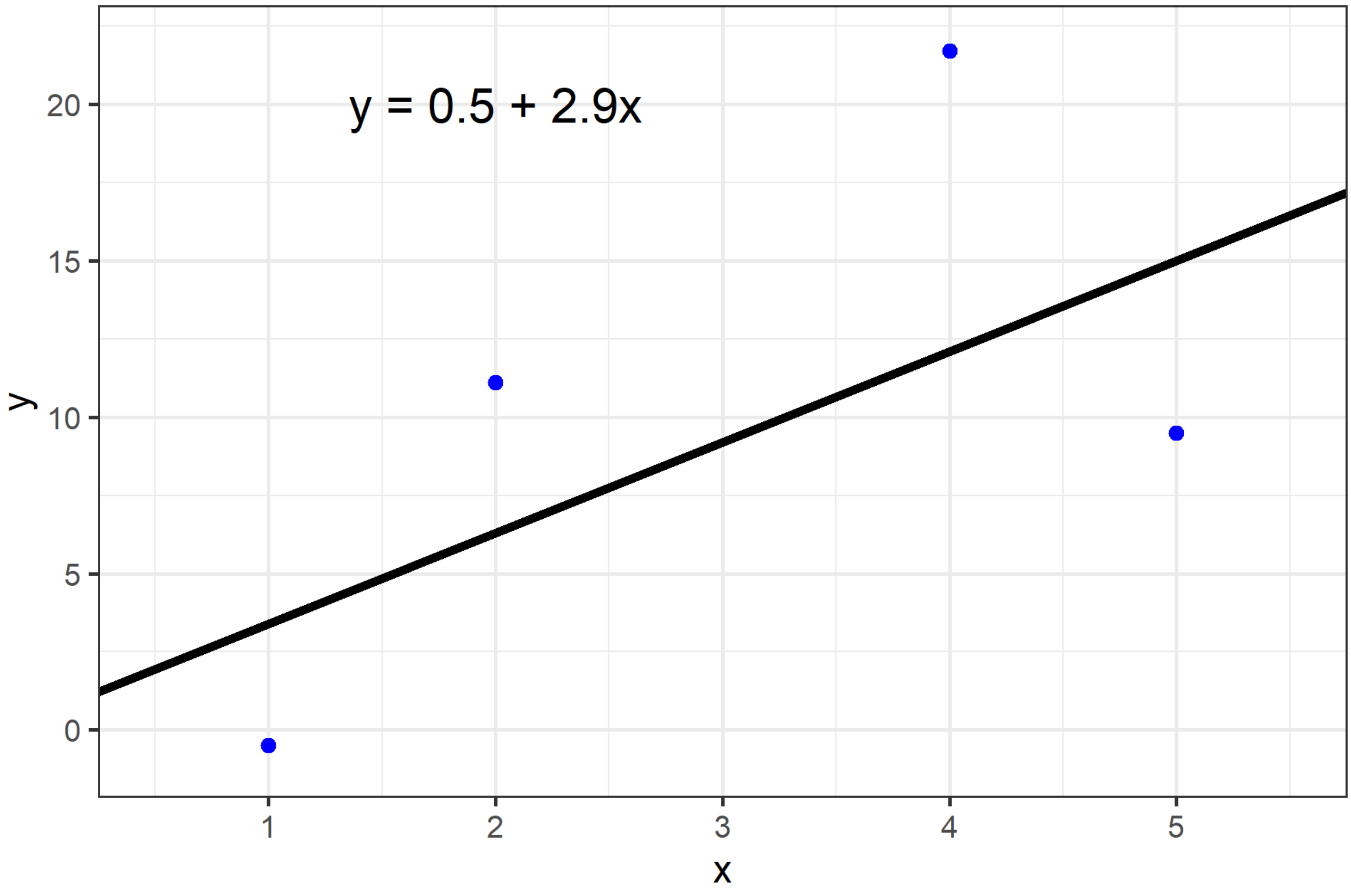
Regressão linear



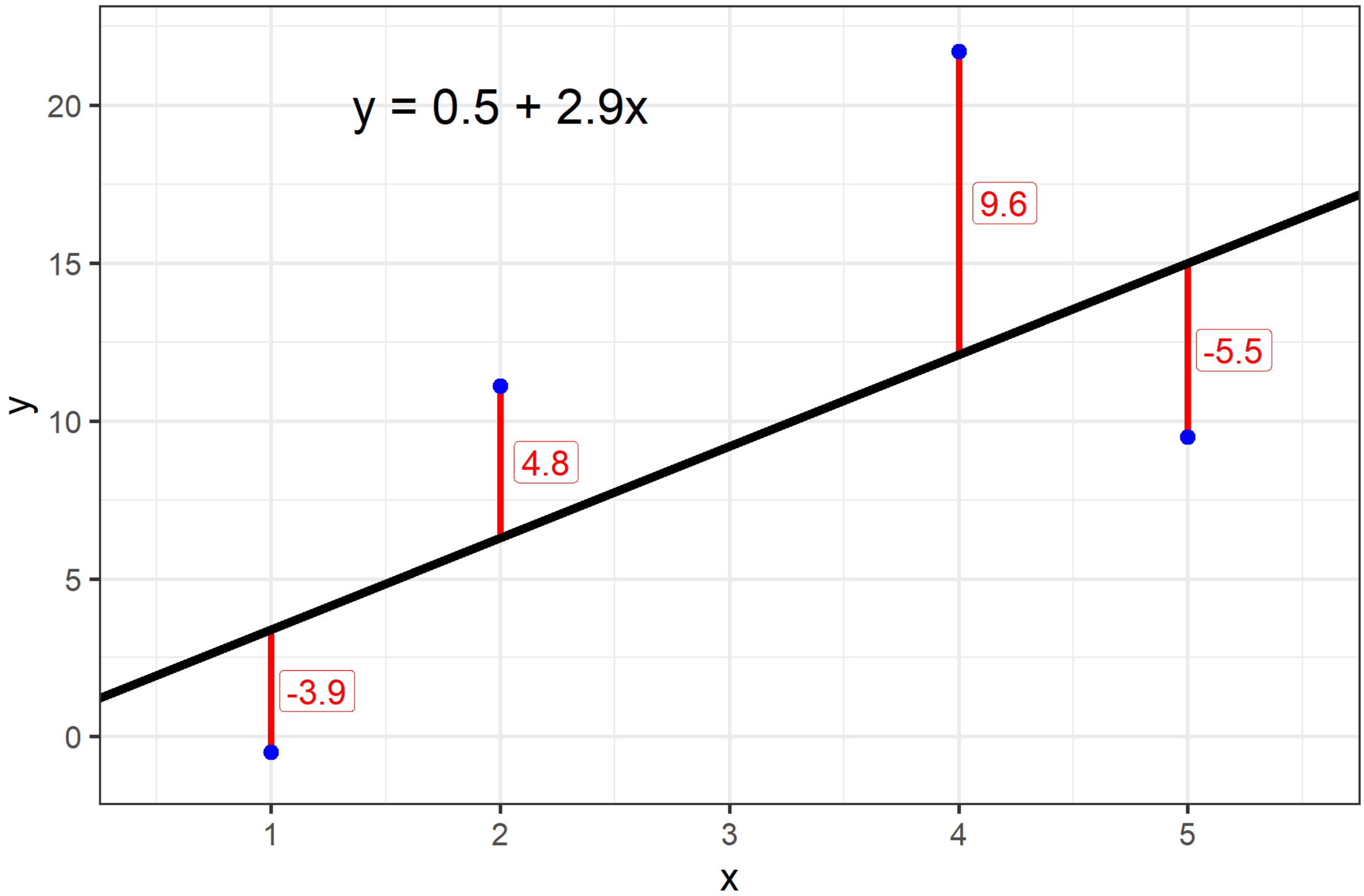
Regressão linear



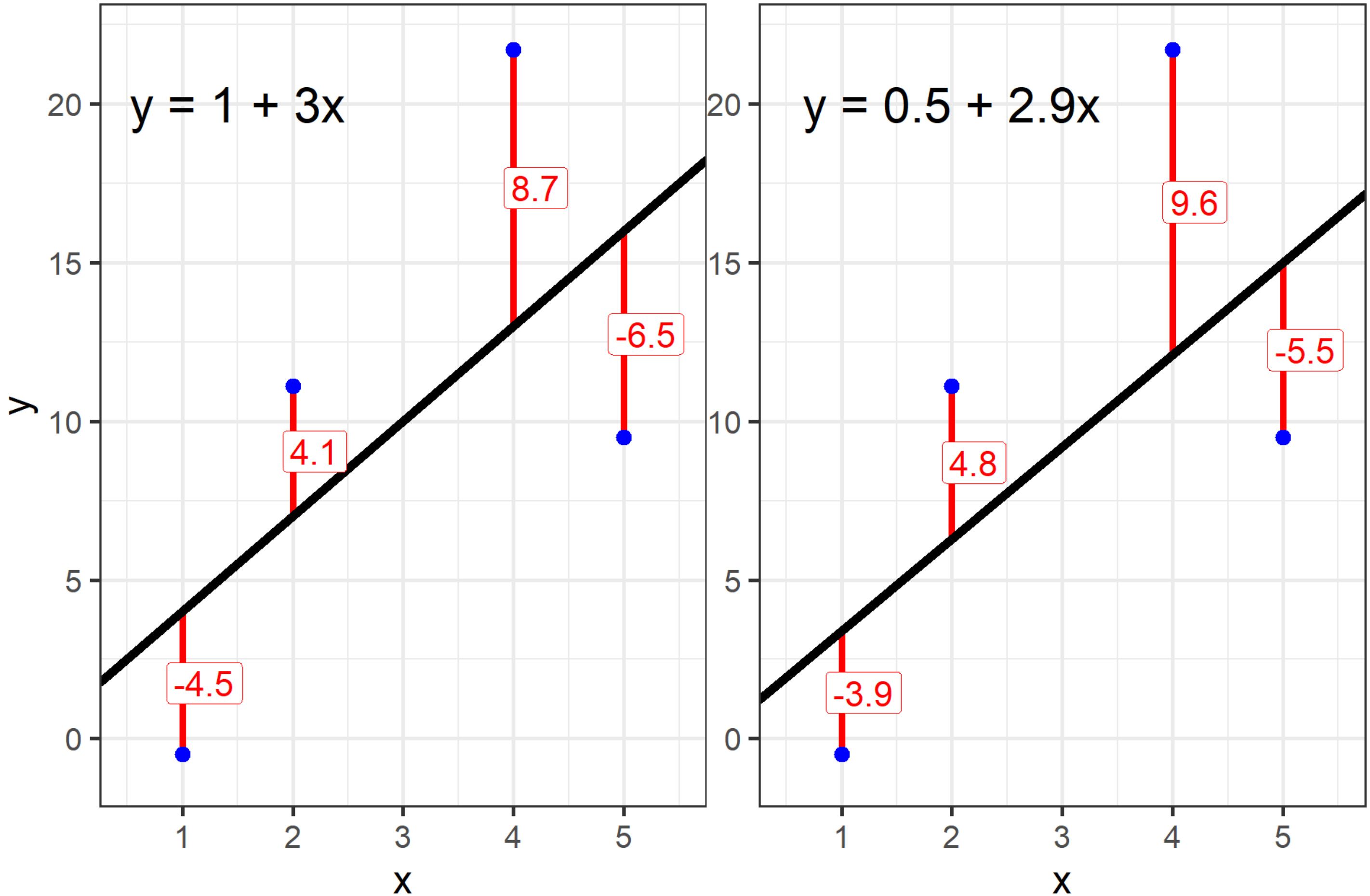
Regressão linear



Regressão linear



Regressão linear



Resíduo

- Denominamos as diferenças $y_i - \hat{y}_i$ de **resíduos**, representados por e_i.
- Como $\hat{y}_i$ é a nossa estimativa para y_i, temos

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

- A soma de quadrados dos resíduos (*residual sum of squares - RSS*) é definida por

$$\begin{aligned} \text{RSS} &= e_1^2 + \cdots e_n^2 \\ &= \left[y_1 - (\hat{\beta}_0 + \hat{\beta}_1 x_1)\right]^2 + \cdots \left[y_n - (\hat{\beta}_0 + \hat{\beta}_1 x_n)\right]^2 \\ &= \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right]^2. \end{aligned}$$

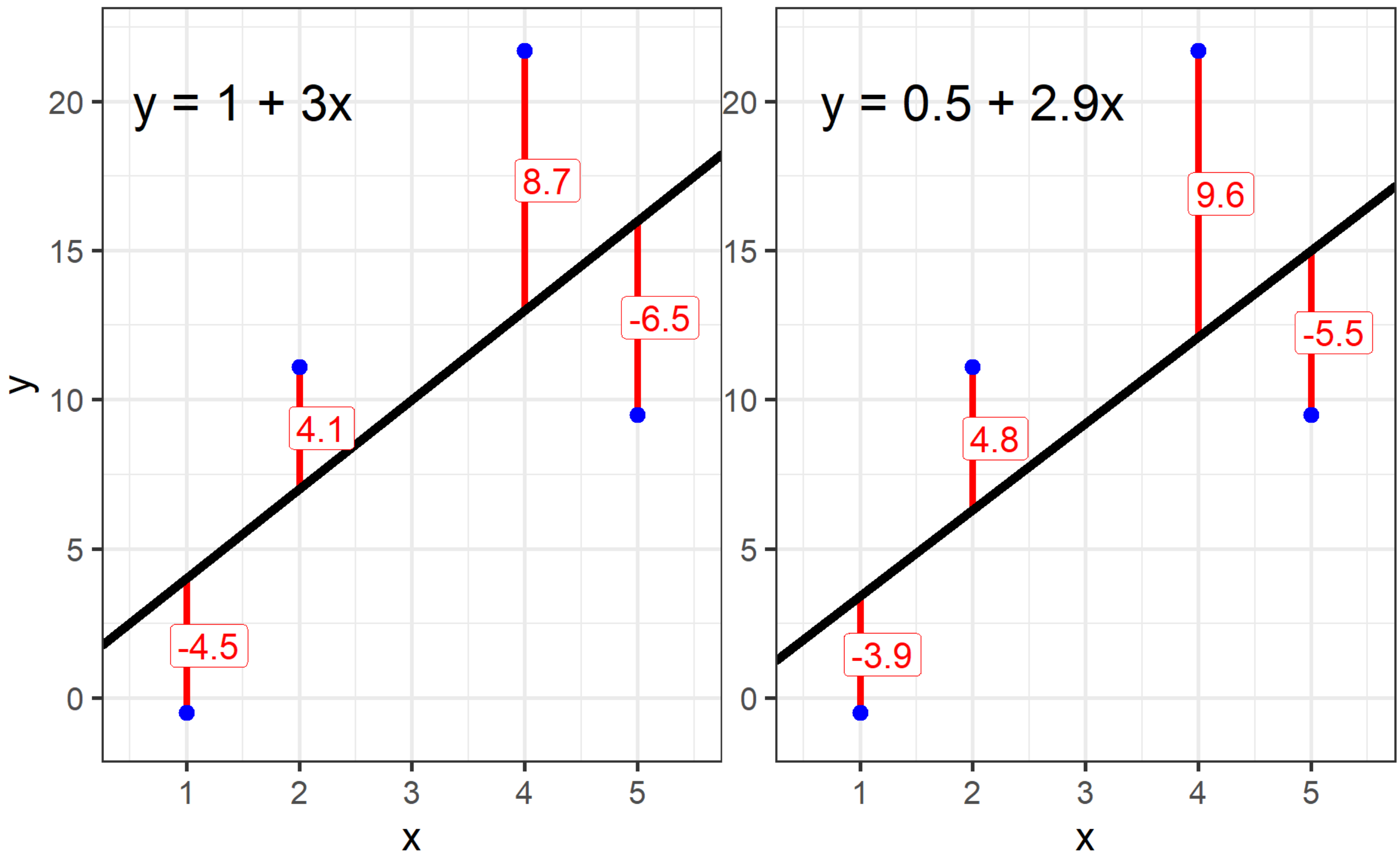
Como estimar os coeficientes?

- Método dos mínimos quaudrados.
- Objetivo: encontrar coeficientes β_0 e β_1 tais que RSS é o menor valor possível.
- Ao minimizar RSS, chegamos em

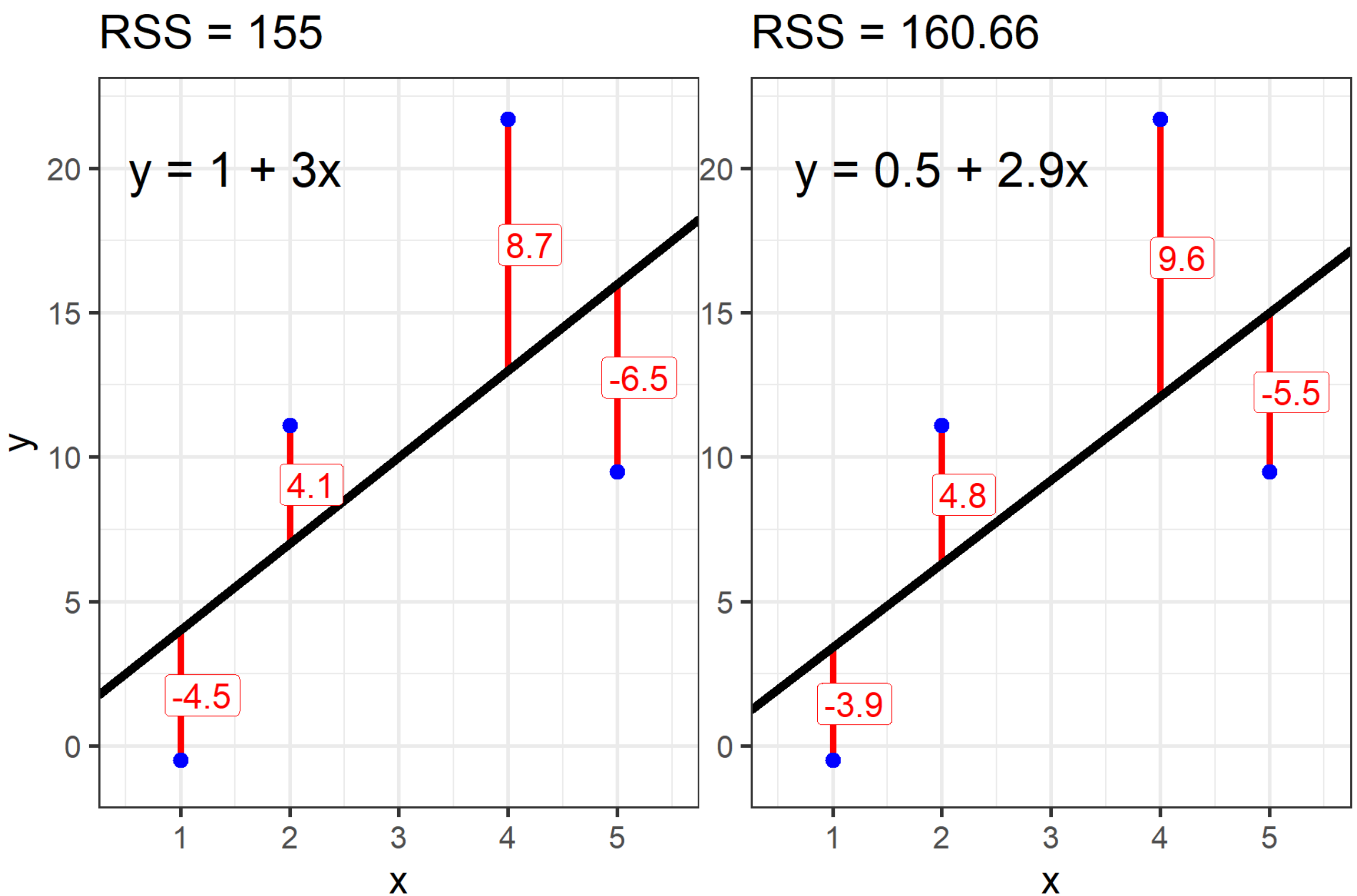
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$

em que $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ e $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Regressão linear



Regressão linear



Qual é o melhor modelo?

Regressão linear

```
##
## Call:
## lm(formula = y ~ x, data = df)
##
## Residuals:
##      1      2      3      4
## -4.83  3.71  8.19 -7.07
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.270      9.415   0.135    0.905
## x              3.060      2.776   1.102    0.385
##
## Residual standard error: 8.779 on 2 degrees of freedom
## Multiple R-squared:  0.3779,	Adjusted R-squared:  0.06683
## F-statistic: 1.215 on 1 and 2 DF,  p-value: 0.3853
```

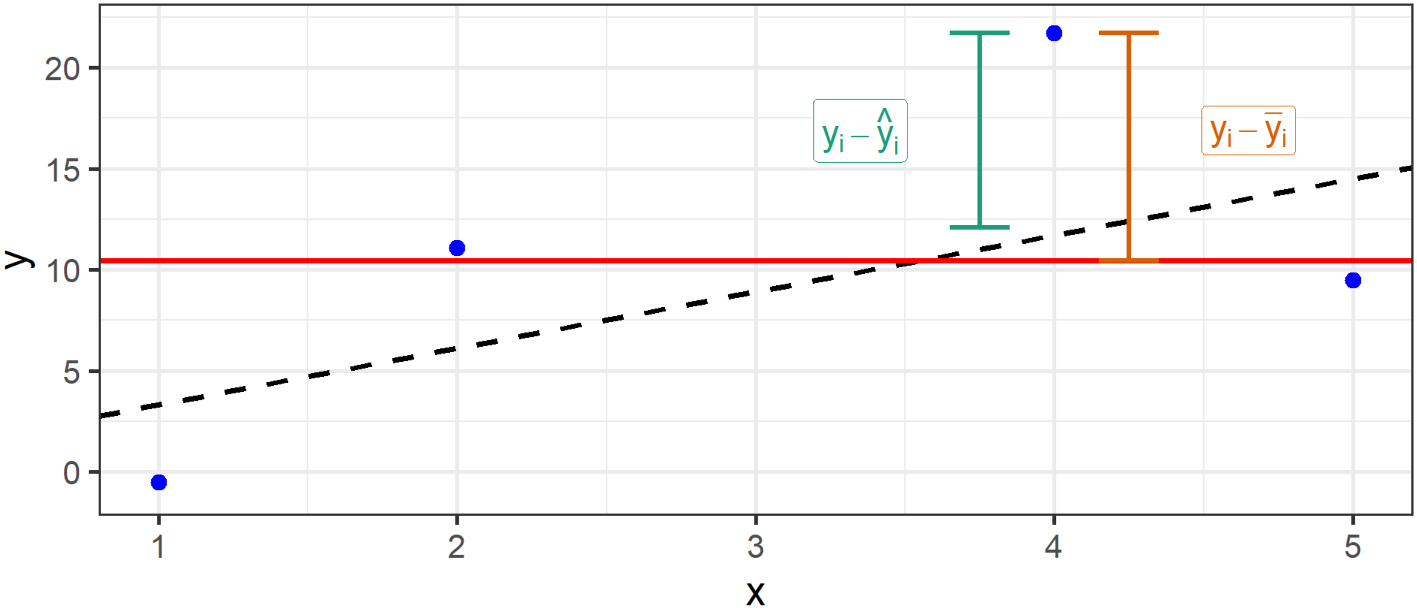
Coeficiente de determinação - R^2

Considere o modelo nulo: $\bar{Y}_i = \frac{1}{n}\sum Y_i$.

O coeficiente de determinação é definido por

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}},$$

em que TSS $= \sum(y_i - \bar{y})^2$ e RSS $= \sum(y_i - \hat{y})^2$.



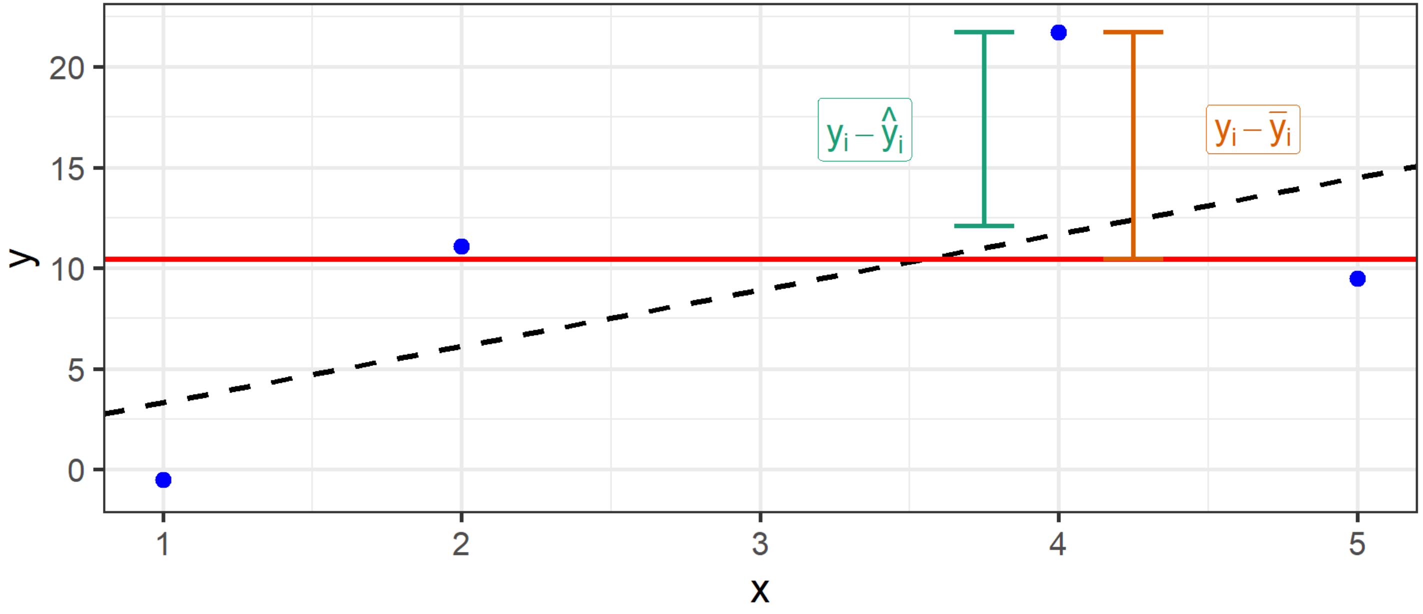
Coeficiente de determinação - R^2 ajustado

Considere o modelo nulo: $\bar{Y}_i = \frac{1}{n} \sum Y_i$.

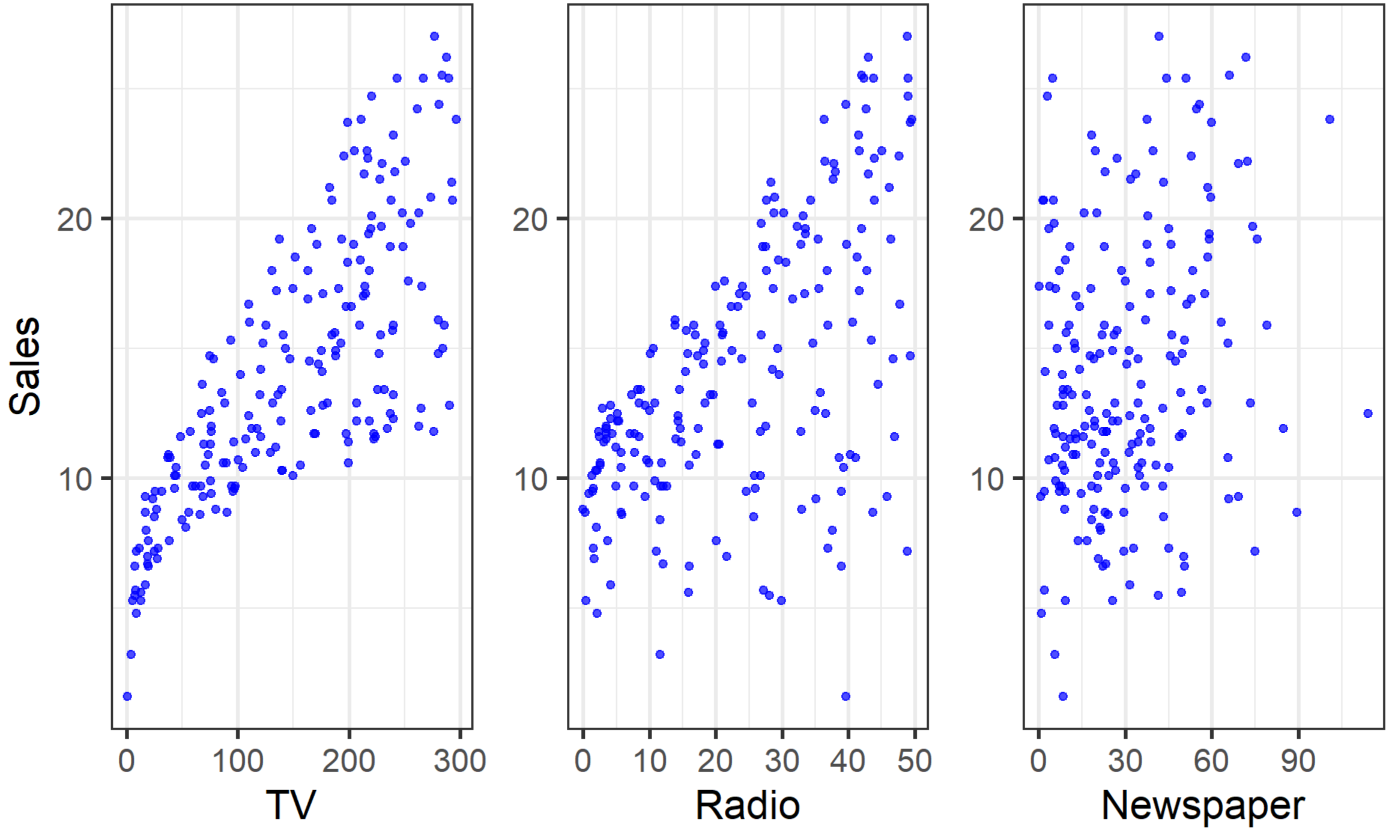
O coeficiente de determinação é definido por

$$R^2_{adj} = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)},$$

em que $\text{TSS} = \sum(y_i - \bar{y})^2$, $\text{RSS} = \sum(y_i - \hat{y})^2$ e d é o número de variáveis no modelo.

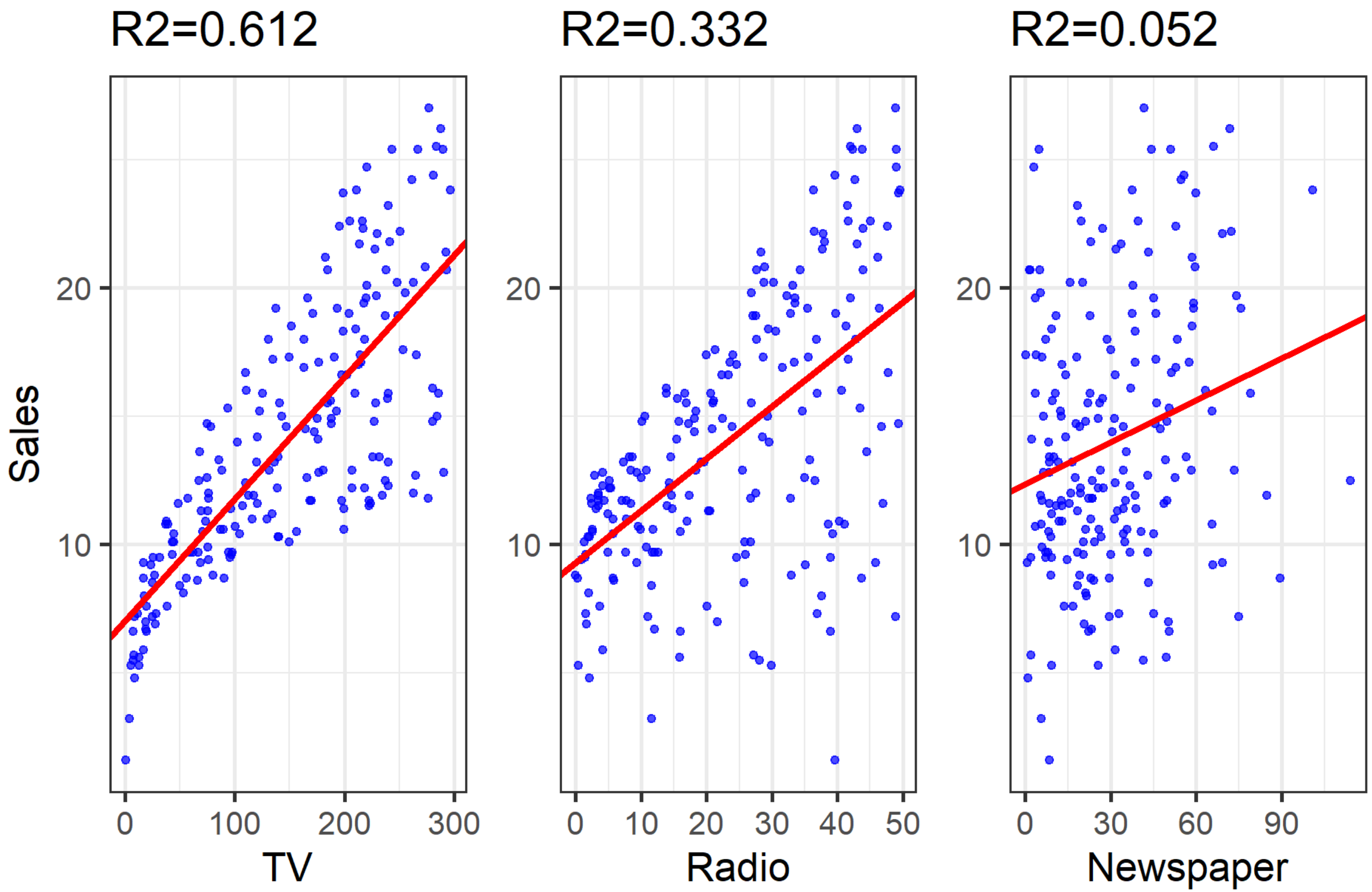


Advertising [1]



[1] Fonte: livro An Introduction to Statistical Learning with Applications in R.

Advertising [1]



[1] Fonte: livro An Introduction to Statistical Learning with Applications in R.

Advertising - R^2

Para o modelo regressão que considera apenas a variável `TV`, temos

$$\text{Sales}_i = \beta_0 + \beta_1 \times \text{TV}_i + \epsilon_i$$

Os coeficientes podem ser estimados através da função `lm`.

```
fit1 <- lm(sales ~ TV, data = advertising)

y_pred <- predict(fit1, advertising)
y_bar  <- mean(advertising$sales)

RSS <- sum((advertising$sales - y_pred)^2)
TSS <- sum((advertising$sales - y_bar)^2)

1 - RSS/TSS
```

```
## [1] 0.6118751
```

```
summary(fit1)$r.squared
```

```
## [1] 0.6118751
```

Advertising - R^2 ajustado

```
y_pred <- predict(fit1, advertising)
y_bar  <- mean(advertising$sales)

RSS <- sum((advertising$sales - y_pred)^2)
TSS <- sum((advertising$sales - y_bar)^2)

1 - (RSS/(nrow(advertising) - 1 - 1))/(TSS/(nrow(advertising) - 1))
```

```
## [1] 0.6099148
```

```
summary(fit1)$adj.r.squared
```

```
## [1] 0.6099148
```

Regressão Linear Múltipla

Em situações práticas, temos $p > 1$ variáveis disponíveis para incluir no modelo, que pode ser escrito como

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

No caso Advertising, temos que $p = 3$. Portanto

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,$$

em que X_1, X_2 e X_3 representam o total investido em TV, rádio e jornal, respectivamente.

Advertising

De acordo com o modelo estimado, temos que

$$\hat{Y} = 2.94 + 0.46X_1 + 0.19X_2 - 0.001X_3.$$

Essas estimativas podem ser intepretadas como

- 2.94 é o valor esperado de vendas associado à quando não é investido nada em publicidade;
- 0.46 é o aumento esperado nas vendas associado à investir uma unidade monetária a mais em publicidade em TV.
- 0.19 é o aumento esperado nas vendas associado à investir uma unidade monetária a mais em publicidade em radio.
- 0.001 é a diminuição esperada nas vendas associada à investir uma unidade monetária a mais em publicidade em jornal.
- **Atenção:** se o objetivo for da análise *inferencial*, testes de hipósteses sobre a significância dos coeficientes β_j devem ser realizados.

Advertising

```
fit <- lm(sales~., advertising)

summary(fit)
```

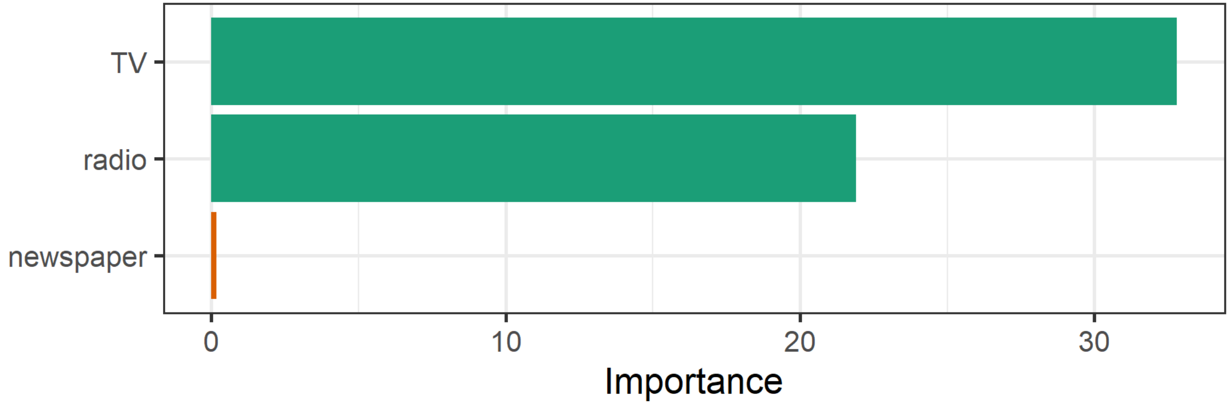
```
##
## Call:
## lm(formula = sales ~ ., data = advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972,    Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

Importância das variáveis

```
library(vip)

vi(fit)
```

```
## # A tibble: 3 x 3
##   Variable  Importance Sign
##   <chr>          <dbl> <chr>
## 1 TV            32.8   POS
## 2 radio         21.9   POS
## 3 newspaper      0.177 NEG
```

```
vip(fit, mapping = aes(fill=Sign))
```



Resumindo

- Definição de métodos para avaliar a performance de um modelo através da estimação do erro de predição:
  - separação em treino e teste;
  - *leave-one-out cross-validation*;
  - validação cruzada em k lotes.
- Modelo KNN para regressão.
- Definição do modelo de regressão linear: $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij}$.
- Estimação dos coeficientes do modelo linear através da minimização do $RSS = \sum_{i=1}^{n} e_i$.
- Como interpretar os coeficientes do modelo de regressão linear.

Obrigado!

magnotfs@insper.edu.br